ISPhS
International Society of Phonetic Sciences

President: Ruth Huntley Bahr

Secretary General: Mária Gósy

Honorary President: Harry Hollien

Vice Presidents:
Angelika Braun
Marie Dohalská-Zichová
Mária Gósy
Damir Horga
Eric Keller
Heinrich Kelz
Stephen Lambacher
Asher Laufer
Judith Rosenhouse

Past Presidents:
Jens-Peter Köster
Harry Hollien
William A. Sakow †
Martin Kloster-Jensen†
Milan Romportl †
Bertil Malmberg †
Eberhard Zwirner †
Daniel Jones †

Honorary Vice Presidents:
A. Abramson      P. Janota      H. Morioka      M. Shirt
S. Agrawal       W. Jassem      R. Nasr         E. Stock
L. Bondarko      M. Kohno       T. Nikolayeva   M. Tatham
E. Emerit        E.-M. Krech     R. K. Potapova  F. Weingartner
G. Fant          A. Marchal      M. Rossi        R. Weiss

Auditor: Angelika Braun

Treasurer: Ruth Huntley Bahr

Affiliated Members (Associations):
American Association of Phonetic Sciences  J. Hoit & W.S. Brown
Dutch Society of Phonetics                  B. Schouten
International Association for Forensic Phonetics  A. Braun
and Acoustics                               
Phonetic Society of Japan                   I. Oshima & K. Maekawa
Polish Phonetics Association                G. Demenko

Affiliated Members (Institutes and Companies):
KayPENTAX, Lincoln Park, NJ, USA            J. Crump
Inst. for Advanced Study of the Communication Processes, University of Florida, USA H. Hollien
Dept. of Phonetics, University of Trier, Germany J.-P. Köster
Dept. of Phonetics, University of Helsinki, Finland A. livonen
Dept. of Phonetics, University of Zürich, Switzerland S. Schmid
Centre of Poetics and Phonetics, University of Geneva, Switzerland S. Vater
President:
Professor Ruth Huntley Bahr, Ph.D.
President’s Office:
Dept. of Communication Sciences
and Disorders
University of South Florida
4202 E. Fowler Ave., PCD 1017
Tampa, FL 33620-8200
USA
Tel.: ++1-813-974-3182
Fax: ++1-813-974-0822
e-mail: rbahr@usf.edu

Secretary General:
Prof. Dr. Mária Gósy
Secretary General’s Office:
Kempelen Farkas Speech
Research Laboratory
Hungarian Academy of Sciences
Benczúr u. 33
H-1068 Budapest
Hungary
++36 (1) 321-4830 ext. 172
++36 (1) 322-9297
e-mail: gosy@nytud.hu

Guest Editors:
Ferenc Bunta, PhD
Guest Editor’s Office:
The University of Houston
Department of Communication Sciences
and Disorders
100 Clinical Research Services
Houston, TX 77204
Tel.: ++1-713-743-2892
e-mail: fbunta@uh.edu

Review Editor:
Prof. Judith Rosenhouse, Ph.D.
Review Editor’s Office:
Swantech
89 Hagalil Street
Haifa 32684
Israel
Tel.: ++972-4-8235546
Fax: ++972-4-8327399
e-mail: swantech@013.net.il

Judit Bóna, PhD
Guest Editor’s Office:
Eötvös Loránd University
Department of Phonetics
Múzeum krt. 4/A
H-1088 Budapest, Hungary
Tel.: ++36 (1) 485-5200 ext. 5189
Fax: ++36 (1) 485-5222
e-mail: bona.judit@btk.elte.hu
The Phonetician
A Publication of ISPhS/International Society of Phonetic Sciences
ISSN 0741-6164
Numbers 103/104 / 2011-I/II

Contents

Editors of this issue

Papers

Frame/content theory as an account of early phonological acquisition by David Ingram, and Kelly Ingram

Coping with the Effects of Coarticulation: Spanish and Portuguese Listeners’ Perception of Nasal Vowel Height by C. Elizabeth Goodin-Mayeda

The production of post-alveolar voiceless fricatives and affricates by bilingual Spanish- and English-speaking children and their monolingual peers: A pilot study by Ferenc Bunta, Christina DiLuca, and Lee Branum-Martin

A case study on the effect of discourse type on fluency level in stuttering by Mária Gósy and Judit Bóna

Automatic evaluation of dysarthric speech and telemedical use in the therapy by Elmar Nöth, Andreas Maier, Arnd Gebhard, Tobias Bocklet, Wilfried Schupp, Maria Schuster, and Tino Haderlein

Ph.D. research

Dissertation research: Strategies of disfluency repairs in spontaneous speech by Dorottya Gyarmathy

Research project

International collaboration between the Department of Phonetics/Phonetics Institute of ELTE and the Bilingual and Cross-Linguistic Language Laboratory at the University of Houston by Ferenc Bunta

Conference Reports

17th International Congress of Phonetic Sciences Hong Kong, China, 17–21 August 2011 by Tekla Etelka Gráczi
12th Annual Conference of the International Speech Communication Association INTERSPEECH 2011 by András Beke ........................................ 100

Book reviews .................................................................................................................................................................. 101

Lingua Americana July-December 2010 year XIV, no. 27, and Lingua Americana January-June 2011, year XV, No. 28. Reviewed by Judith Rosenhouse ........................................................................................................ 101


Meetings, conferences and workshops ...................................................................................................................... 112

Call for Papers .............................................................................................................................................................. 117

ISPhS Membership Application Form ............................................................................................................................. 118

News on Dues ................................................................................................................................................................. 119
EDITORS OF THIS ISSUE

As the President of ISPhS, it is with great pleasure and gratitude that I present the two editors of this issue to you. They have done an excellent job!

Ferenc Bunta is an assistant professor in the Department of Communication Sciences and Disorders at the University of Houston. His research focuses on bilingual and cross-linguistic phonological acquisition. Dr. Bunta received his Ph.D. from Arizona State University, Department of Speech and Hearing Science and completed a postdoctoral fellowship in the Department of Communication Sciences and Disorders at Temple University with a joint appointment in the Bilingual Language Laboratory and the Eleanor M. Saffran Center for Cognitive Neuroscience before joining faculty at the University of Houston. Dr. Bunta has taught courses on bilingualism, phonetics, phonology, writing in communication sciences and disorders, speech and language acquisition, and speech science.

Judit Bóna holds a Master's degree in Hungarian language and literature and Italian language and literature from the Eötvös Loránd University, Budapest, Hungary. She earned her PhD degree in Phonetics in 2007 at the Eötvös Loránd University in Budapest, Hungary. Since 2002, she has been teaching courses in Phonetics and Psycholinguistics at the Department of Phonetics of Eötvös Loránd University. The main focus of her research is the temporal characteristics of speech, disfluencies and self-corrections in speech, speech production in the elderly, and fluency disorders.
Abstract

Frame/Content Theory (Davis, MacNeilage, & Matyear, 2002) claims that children’s early words are constrained by basic mandibular movements associated with chewing and sucking, with no influence of perception or phonological representations. The present study conducted phonetic and phonological analyses of the first words of two children between 13 and 23 months of age. The results only partially supported Frame/Content Theory, and found evidence of the influence of perception and phonological organization. It is concluded that Frame/Content Theory is too narrow in its focus on token frequencies and did not capture the full range of patterns in the children’s words.

1 Introduction

There has been a long standing debate about whether children begin phonological acquisition at the onset of word acquisition (Jakobson, 1968), or at a later point (e.g. after 50 or more, Locke, 1983) once articulatory restrictions due to the maturing speech tract begin to lessen. A series of publications by MacNeilage and Davis (e.g. Davis & MacNeilage, 1995; Davis, MacNeilage & Matyear, 2002; Davis & Zajdo, 2008; MacNeilage, 1998; MacNeilage & Davis, 2000, 2001) have argued for the latter perspective, proposing an approach called Frame/Content Theory (FCT). FCT claims that early word productions, in particular, the first syllables, are constrained by basic mandibular movements associated with chewing and sucking. These movements optimally produce three basic CV syllable types, ‘di’, ‘ba’, and ‘gu’. Stated more generally, coronal tongue movements are associated with front vowels, labial productions with central vowels, and dorsal tongue movements with back (rounded) vowels. Evidence in support of FCT is given by analyses of early consonant-vowel interactions from the babbling of six English infants, the first words of 10 English children, and additional analyses of similar data from children acquiring French, Swedish, Japanese, Ecuadorian-Quichua, and Brazilian-Portuguese (cf. review in MacNeilage & Davis, 2000). Partially supportive data of FCT for Cantonese is reported in Chen and Kent (2005).

The general patterns predicted by FCT can be seen by examining the data from the first words of 10 English children reported in the Appendix of Davis, MacNeilage & Matyear (2002). The children ranged in age from 1;3 to 3;0 (mean = 1;11) and had vocabularies ranging from 77 to 398 words (mean = 148). The consonant-vowel interactions are measured by ratios comparing the actual distributions of three categories of consonants (coronal, labial, dorsal), with three categories of vowels (front, central, back), as compared with expected ratios, given the frequencies of the six categories. The expected ratios should be 1.00. The ratios are determined as
follows. Suppose the sample consisted of 48 words, and 12 words had labial consonants (0.25 of all words), and 24 words had front vowels (0.5 of all words). The expected number of words with labial consonants occurring with front vowels would be 6 (0.25 times 0.5, or 0.125 of all words). If 8 words actually had this combination, the ratio of observed to expected would be 1.33, or above 1.0. If the actual occurrences were lower, however, say 4 instances, then the ratio would be 0.67. Table 1 is an example of the results from their subject C based on such ratios.

Table 1. Ratios of observed to expected occurrences for child C (Davis et al. 2002)

<table>
<thead>
<tr>
<th>Consonants</th>
<th>Front</th>
<th>Central</th>
<th>Back</th>
</tr>
</thead>
<tbody>
<tr>
<td>Coronal</td>
<td>1.26</td>
<td>0.77</td>
<td>0.92</td>
</tr>
<tr>
<td>Labial</td>
<td>0.93</td>
<td>1.11</td>
<td>0.91</td>
</tr>
<tr>
<td>Dorsal</td>
<td>0.68</td>
<td>1.00</td>
<td>1.54</td>
</tr>
</tbody>
</table>

The ratios in bold indicate higher occurrences for these interactions than to be expected from the frequency of the individual categories, and thus support for FCT. The overall ratios across the 10 children were 1.48 for coronal-front, 1.29 for labial-central, and 1.39 for dorsal-back. It should also be noted this effect extends well into the early months of word learning, since the children had a mean age of 23 months and some vocabularies up to nearly four hundred words.

While the results in these studies support Frame/Content Theory, there are limitations to this line of research that require further inquiry. First and foremost, the analyses conducted focus solely on token data, i.e., all child productions. The mean number of tokens in Davis et al. (2002) was 842, over five times greater than the children’s lexical types (i.e. words). As argued in Newmeyer (2003), there needs to be a difference drawn between usage and knowledge. Statistical tendencies in speech do not necessarily reflect in all instances underlying linguistic (or in the current case, phonological) knowledge. The fact that a child may prefer to produce more frequently ‘di’ syllables to ‘du’ ones, for example, does not mean that he or she will not use ‘du’ syllables as phonological forms in their system. A phoneme may be restricted in its environments, and possibly only appear in low frequency words, but it would be considered an acquired phoneme nonetheless. A similar argument can be made considering the perceptual discrimination of speech sounds. As shown in Miller and Nicely (1955), English speakers show degrees of difficulty in perceiving certain phonemes over others, for example that /l/ is easier to perceive than /θ/.

These difficulties, however, do not negate the fact the /l/ and /θ/ are English phonemes.

For FCT to meet the claims made, it should be equally capable of capturing children’s word types as well as word tokens. This point can be demonstrated by examining a four minute sample from a child in the first author’s database on early speech acquisition. The child was a boy, D, who was 1;6.25 at the time of the sample, with an expressive vocabulary of only six words. During the sample, D produced 30 vocalizations, a mix of babbling with a few intelligible words. There were a total of 57 syllables with CV interactions. An FCT analysis of the 57 syllables supported the predicted pattern given in Table 1. D’s observed-to-expected ratios for
coronal-front, labial-central, and dorsal-back interactions were 2.11, 1.79, and 1.64 respectively. D only produced three words in the sample, however, *byebye* [baba], *this* [di:], and *cookie* [gu], [guig]. These words show the predicted syllables (except for [gi]). The point here is that FCT should be able to account for the word types and well as the phonetic tokens as D’s vocabulary increases.

The present study is an attempt to assess Frame/Content Theory by applying it in two case studies of early phonological acquisition, and expanding the analyses to explore the hypothesis through an analysis of the 4 phonological aspects just discussed. It is hypothesized that FCT should also be capable of capturing a child’s type frequencies, as well as token frequencies. That is, it should not only predict that a child prefers to produce ‘di’, ‘ba’ and ‘gu’ syllables more frequently, but that the child prefers to use these for more distinct words than other syllables. It is also hypothesized that a preference for these syllables should be found in the analysis of a child’s homonyms. Under our interpretation of FCT, it should predict that instances of homonymy will be greater for preferred syllables than for those that are less preferred. The children’s data are also analyzed for evidence of phonological organization. FCT predicts that such evidence should be lacking in three ways. There should be phonological gaps of non-preferred syllables, no evidence of perceptual factors playing a role in the child’s word productions, and no instances of unique phonological forms.

It should be emphasized that we do not challenge the results of FCT regarding children’s babbling, nor the distribution of the preferred syllables in children’s word tokens. The issue at stake is the conclusion in Davis et al. (2002) that FCT is superior to phonological accounts of early speech acquisition. The present study provides the phonological analyses not found in the FCT studies in order to provide the data necessary to make such a broadly based conclusion.

2 Method

2.1 Participants

The first child studied, Samuel, is the first-born son of the authors, and a monolingual child acquiring English. A diary was kept of his word productions between 1;3 and 1;11, with a total of 108 words recorded. A conservative approach was taken regarding the selection of words, since the purpose of the study was to determine the predictions of FCT for the child’s emerging phonology. We wanted to increase the possibility that the analyses conducted were on established words, not emerging or questionable ones. A production was considered a word if it was an identifiable word in English, allowing leniency for child forms such as *baa*, the sound that a sheep makes, which Samuel used for labeling sheep. Words were only entered for analysis if both parents had heard Samuel produce the word, and agreed on the transcription. This procedure eliminated potentially emerging words of low frequency or imitations. Samuel was slow in his phonological acquisition, and showed a highly restricted system of syllables, consonants, and vowels.

The second child studied, Rachel, a girl, is the second-born child of the authors, and a monolingual child acquiring English. (Samuel was age 2;1 at the time of Rachel’s birth). A diary was kept of her word productions between 1;1 and 1;8, following the same procedure as used with Samuel, for a total of 72 words recorded. Rachel was also slow in her phonological acquisition, showing a highly restricted system of syllables, consonants, and vowels.
Both children’s ages are in the lower range of the children studied in Davis et al. (2002) (Davis et al. mean age 1;11), with their vocabulary falling on the low side (Davis et al. range 76 to 398 words). Both children were also slow in their phonological acquisition. The children, therefore, should be optimal for demonstrating the predictions of FCT.

2.2 Analyses

Analyses were conducted that tested the predictions of FCT versus those for phonological organization. First, syllables used in all phonetic types (i.e. unique forms for each target word), were categorized into one of the nine types used in Table 1, e.g. coronal-front, coronal-central, etc.). As was done in Davis et al. (2002), the vowel [a] was scored as a central vowel, and [w] as a labial consonant. Words were excluded from the analysis when they contained no syllables with consonant-vowel interactions, e.g. flower [o], or had [h] as an onset. Syllables with the diphthongs [au], [ai], and [ai] were excluded because it was not clear what predictions FCT makes about complex vowels with parts falling into different categories.

Once the actual distributions of all the possible interactions were determined, ratios of predicted versus actual occurrences of consonant-vowel interactions were calculated and placed in chart form as Table 1.

Homonyms were defined as words that had a phonetic form that was the same as at least one other word, e.g. pool, Pooh [bu]. Syllables in homonyms with two syllables were each scored separately, e.g. bubble, pretzel [bubu]. The actual frequencies of syllables in homonyms only were determined as was done above for all words. Ratios of predicted versus actual occurrences were then calculated.

Phonological analyses looked at the capacity for children’s consonants to occur with their vowels across word types, using the frequency criterion that only consonants that occurred in at least five words were analyzed. This frequency criterion was used since the vowels of each child’s words were placed into five categories, (i.e. [ i, a/e, a/A, o/o, u]). If a consonant was used in fewer than five words, e.g. [l], it would not be possible to say with certainty whether its occurrence was due to FCT or to a sampling error.

Lastly, individual lexical items were examined to find evidence of phonological organization based on the perceived form of words, or due to unique organization of the emerging sound system. For example, the word Murphy could be produced either as [ma], a syllable consistent with Samuel’s phonetic inventory and with FCT, or as [mi], a syllable not predicted by FCT but one that would reflect an effort to produce segments from both syllables of the target word. Unique productions would be words that consisted of sounds from the child’s system, but not ones meeting the general patterns of other words. For example, Samuel acquired the word Kip, the name of one of the family dogs. Based on similar words, we would anticipate that he would produce it as [gi], his form for cheese and the letter G, or [gap], a form he used as a template for words with dorsal stops in the target, or possibly [gu] or [di], forms predicted by FCT. Instead, he produced it as [dædiba], a unique production for just this word.
3 Results
3.1 Results for Samuel

Of the 108 words studied, sixty percent were CVs, with CV and CVCV syllable shapes combining to capture 82% of the total words. Samuel’s productions were by and large limited to six consonants [m, n, b, d, g, w], and five categories of vowels [i, æ, a, o, u]. Certain vowels varied and are thus shown as a single vocalic category. These categories conform to the ones used in FCT, i.e. front, central and back vowels. His vocabulary also contained a high percentage of homonyms, with 61% of his words homonymous with at least one other word.

Samuel’s frequencies and ratios of observed to expected occurrences of consonants and vowels, based on the child’s word types, were as shown in Table 2. Those interactions predicted by FCT to be greater than expected are shown in bold.

Table 2. Frequency and ratios of observed to expected occurrences for Samuel

<table>
<thead>
<tr>
<th></th>
<th>Observed</th>
<th>Observed/expected ratios</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Vowels</td>
<td>Vowels</td>
</tr>
<tr>
<td></td>
<td>Front</td>
<td>Central</td>
</tr>
<tr>
<td>Coronal</td>
<td>27</td>
<td>4</td>
</tr>
<tr>
<td>Labial</td>
<td>30</td>
<td>17</td>
</tr>
<tr>
<td>Dorsal</td>
<td>4</td>
<td>7</td>
</tr>
</tbody>
</table>

χ² (4, N=9) = 13.2, p = .01

The chi-squared analysis, the primary statistic used by Davis et al. (2002), reached significance, indicating that there were consonant-vowel interactions, though the small number of words requires this result to be interpreted with extreme caution. The observed to expected ratios provided partial support for FCT. The interaction between coronals and front vowels was as the theory predicts, and had a major effect on the variance in the chi-squared analysis, along with the low occurrence of the coronal-central interaction. The data supported the interaction for labials in that they occurred at a greater than expected rate with central vowels. Dorsal consonants did not follow the predicted pattern in that they occurred at greater than expected rates with central vowels than with back vowels. This result, however, was due to a very high rate of occurrence of the phonetic form [gap]. When this form was removed from the analysis, the ratios became 0.41, 0.81, and 2.29 respectively for the dorsal consonants with front, central, and back vowels.

Samuel’s word productions consisted of many homonyms, with over half (61%) being homonymous with at least one other word. The distribution of these by syllable type is given in Table 3 and the ratios of observed to expected co-occurrences is given in Table 4.

The interactions approached but did not reach significance at the .05 level (again with the caution mentioned about this statistic above). Homonyms were found for all nine possible interactions, though certain ones were much more frequent than others, (i.e. labial/front, labial/back, and coronal/front). Only the latter would be predicted by FCT. Most interesting, however, was that the direction of the usage was to use the
preferred syllables less often for homonyms than the non-preferred syllables, that is, the opposite of the prediction made. The results, therefore, could be interpreted as support of FCT if the hypothesis were revised to make the opposite of the prediction made in the introduction. We will return to this point in the Discussion.

**Table 3.** Number of homophonic CV words by consonant vowel interaction for Samuel

<table>
<thead>
<tr>
<th>Consonants</th>
<th>Front</th>
<th>Central</th>
<th>Back</th>
<th>CV Forms</th>
</tr>
</thead>
<tbody>
<tr>
<td>Coronal</td>
<td>8</td>
<td>3</td>
<td>4</td>
<td>næ (2), di (6), næ (3), du (3)</td>
</tr>
<tr>
<td>Labial</td>
<td>20</td>
<td>5</td>
<td>17</td>
<td>bæ (5), bi (9), bæbi (2), mâe (2), wewe (2) bæba (2), bæbu (3) bu (9), bæ (2), mu (3), babu (3)</td>
</tr>
<tr>
<td>Dorsal</td>
<td>4</td>
<td>5</td>
<td>2</td>
<td>gi (4), gap (5), go (2)</td>
</tr>
</tbody>
</table>

**Table 4.** Ratios of observed to expected occurrences of homonyms for Samuel

<table>
<thead>
<tr>
<th>Consonants</th>
<th>Front</th>
<th>Central</th>
<th>Back</th>
</tr>
</thead>
<tbody>
<tr>
<td>Coronal</td>
<td>0.53</td>
<td>1.24</td>
<td>0.82</td>
</tr>
<tr>
<td>Labial</td>
<td>1.27</td>
<td>0.63</td>
<td>1.15</td>
</tr>
<tr>
<td>Dorsal</td>
<td>1.65</td>
<td>1.37</td>
<td>0.82</td>
</tr>
</tbody>
</table>

χ² (4, N=9) = 8.65, p = .07

The phonological analysis of Samuel’s five most frequent consonants with his five most frequent vowel categories found 27 of the 30 possible combinations (90%). These are shown in Table 5.

**Table 5.** Samuel’s CV interactions with the most common consonants and vowels

<table>
<thead>
<tr>
<th>Vowels</th>
<th>i</th>
<th>æ/ɛ</th>
<th>a/ʌ</th>
<th>æ/o</th>
<th>u</th>
</tr>
</thead>
<tbody>
<tr>
<td>Consonants</td>
<td>[m]</td>
<td>mi</td>
<td>mâe</td>
<td>ma, mâ</td>
<td>mu</td>
</tr>
<tr>
<td></td>
<td>[n]</td>
<td>ni</td>
<td>næ</td>
<td>na</td>
<td>nu</td>
</tr>
<tr>
<td></td>
<td>[b]</td>
<td>bi</td>
<td>bæ</td>
<td>bu</td>
<td>bɔ</td>
</tr>
<tr>
<td></td>
<td>[d]</td>
<td>di</td>
<td>dæ</td>
<td>da</td>
<td>dɔ</td>
</tr>
<tr>
<td></td>
<td>[g]</td>
<td>gi</td>
<td>ga</td>
<td>gɔ, go</td>
<td>gu</td>
</tr>
<tr>
<td></td>
<td>[w]</td>
<td>wi</td>
<td>wæ</td>
<td>wa</td>
<td>wo</td>
</tr>
</tbody>
</table>

Two consonants, [b] and [d], occurred with all five vowel categories. These results indicate that Samuel’s restricted inventories did not stop him from exploring most of
their possible interactions. The three missing syllable types, however, were non-preferred syllables based on FCT.

Lastly, there were several words found that indicated linguistic organization of the input vocabulary, suggestive of phonological representations. Here are four patterns that are not explainable by a theory such as FCT that predicts all child productions on the basis of articulatory factors:

1. One of the family dogs was named Murphy. Given Samuel’s productions, this could be predicted to be [ma] or reduplicated [mamam], using a central vowel. This is also the syllable predicted by FCT. Samuel, however, produced the word as [mi], using the consonant from the first syllable and the vowel from the second; this production suggests an effort to capture sounds from both syllables;

2. A similar adaptation was made for the name Samuel. He was unable to produce fricatives at this stage, and replaced lingual fricatives with [d]; e.g., seed [di], shoe [du], sun [dae]. This pattern along with his other productions would predict any of the following forms, [dae], [daede], [daedu], [mu], [memae], [memmu]. Instead, he produced [menu], using it consistently for several months. This production has the characteristic of maintaining the coronal feature of the first consonant and the nasal feature of the second. Like his word for Murphy, it showed an effort to capture more of the target word than a single syllable. The initial syllable would be predicted by the FCT, but the final one would not;

3. While the majority of Samuel’s words were monosyllables, there were a smaller number of disyllables that were partial or complete reduplications with stress on the first syllable. One such common word was daddy [‘daedi]. At the same time, he had two words with the final syllable stressed. One of these words was happy [dae’di], which would otherwise have been a homonym with daddy. Another similar pair was his production of grandma and [‘mama] vs. mommy [ma.’i]. If his general patterns were applied, these two words would have been homonyms. It has been suggested in Ingram (1975), that children at an early stage of phonological acquisition will make unique adjustments like these to maintain surface contrasts between words of importance that would otherwise become homonyms;

4. Another family dog was named Kip. As discussed earlier, expected productions might be [gi], his form for cheese and the letter G, [g/ap], a form he used as a template for words with dorsal stops in the target, or possibly [gu] or [di], forms predicted by FCT.

Instead, however, Samuel created a proto-word [‘daediba] for Kip. The middle vowel would vary, but the first and last ones were consistently made, and the form lasted for several months, even when the form [gip] began to be used. This is a case of using a more complex form involving three syllables for a simple CV. The origin for the word was a mystery. Another (less mysterious) example of Samuel using a longer form over a simpler one was his production of brown bear [‘bubububu’ba], which varied with [‘bubububu’ba].

3.2 Results for Rachel

Like Samuel, Rachel’s words were primarily CV and CVCV. She was more multisyllabic than Samuel, however, with 50% of her words being CVCV and 43% CV. Her productions were by and large limited to five consonants [m, n, b, d, w], and five categories of vowels [ i, æ/ɛ, θ/ʌ, ɔ/ø, u]. Despite an even more restricted phonological system than Samuel, Rachel’s words showed less homonymy, with 19% being homonymous with at least one other word.
Rachel’s frequencies and ratios of observed to predicted occurrences of consonants and vowels, based on the child’s word types, were as shown in Table 6. Those interactions predicted by FCT to be greater than expected are shown in bold.

**Table 6. Frequency and ratios of observed to expected occurrences for Rachel**

<table>
<thead>
<tr>
<th>Observed/expected ratios</th>
<th>Observed</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vowels</td>
<td>Vowels</td>
</tr>
<tr>
<td>_________________________</td>
<td>_________</td>
</tr>
<tr>
<td>Consonants</td>
<td>Front</td>
</tr>
<tr>
<td>Coronal</td>
<td>29</td>
</tr>
<tr>
<td>Labial</td>
<td>30</td>
</tr>
</tbody>
</table>

χ² (4, N=9) = 2.97, p = .21

The results do not provide much support for the FCT. The occurrence of coronal consonants with front vowels was about what would be expected for the actual frequencies, and the occurrence of coronals with back vowels was actually higher. This result was due to Rachel’s fronting of dorsal consonants, e.g. *goat* [do], *choo choo* [dudu].

As for Labial consonants, there was a slightly higher occurrence with central vowels, but they also occurred about as expected for the actual frequencies with front vowels. Since there were no dorsal consonants, there were no dorsal consonant vowel interactions of any kind. If these are preferred ‘gu’ syllables, we should expect to see them, or to see a lack of back vowels with the other consonants. As just observed, however, the back vowels were the most preferred ones with coronal consonants.

Unlike Samuel, Rachel’s word productions consisted of very few homonyms, with only 19% of her words being homonymous with other words. The distribution of these by syllable type is given in Table 7. The ratios of observed to expected occurrences of homonyms are in Table 8.

These ratios did not come close to significance, and suggest that the few homonyms she used were randomly spread across her syllables. There was no support for the opposite pattern from our prediction, unlike that found for Samuel.

**Table 7. Number of homophonic CV words by consonant vowel interaction for Rachel**

<table>
<thead>
<tr>
<th>Consonants</th>
<th>Front</th>
<th>Central</th>
<th>Back</th>
<th>CV Forms</th>
</tr>
</thead>
<tbody>
<tr>
<td>Coronal</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>dædi (2), da (2), dop (3)</td>
</tr>
<tr>
<td>Labial</td>
<td>5</td>
<td>4</td>
<td>2</td>
<td>bæ (3), mæ (2), ba (2), ma (2), wo (2)</td>
</tr>
</tbody>
</table>

Table 9 gives the occurrence of CV interactions that Rachel used, given her limited system of just 5 consonants and 5 vowels. Of the 25 possible combinations,
22 or 88% were found, a result very similar to those for Samuel. One of the missing syllable types was [ni], a preferred syllable according to FCT.

Table 8. Ratios of observe to expected occurrences of homonyms for Rachel

<table>
<thead>
<tr>
<th>Consonants</th>
<th>Front</th>
<th>Central</th>
<th>Back</th>
</tr>
</thead>
<tbody>
<tr>
<td>Coronal</td>
<td>0.75</td>
<td>1.05</td>
<td>1.05</td>
</tr>
<tr>
<td>Labial</td>
<td>1.20</td>
<td>1.11</td>
<td>0.70</td>
</tr>
</tbody>
</table>

\[ \chi^2 (4, N=9) = 0.7, p = .ns \]

Table 9. Rachel’s CV interactions with the most common consonants and vowels

<table>
<thead>
<tr>
<th>Vowels</th>
<th>Consonants</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>i</td>
</tr>
<tr>
<td>[m]</td>
<td>mi</td>
</tr>
<tr>
<td>[n]</td>
<td>næ</td>
</tr>
<tr>
<td>[b]</td>
<td>bi</td>
</tr>
<tr>
<td>[d]</td>
<td>di</td>
</tr>
<tr>
<td>[w]</td>
<td>wi</td>
</tr>
</tbody>
</table>

Unlike for Samuel, there were few instances of unique productions demonstrating phonological re-organization of word forms. The primary perceptual effect in Rachel’s productions was the preservation of syllables. Rachel was a child who compensated for her limited consonantal accuracy by maintaining syllable shapes, a trade off effect reported years ago in Macken (1978). An example is Rachel’s production Tiki hut as [di di d\|].

4 Discussion

The results provide partial support for the Frame/Content Theory in the form of Samuel’s word types (Table 2). It should be noted that this result was aided by the elimination of a frequent CVC shape [g\|p] for the dorsal-back interaction. This effect represents an inherent difficulty with looking at these interactions in word types in the first one hundred words or so. The numbers are so small that an additional form or two here and there can greatly influence the analysis. The results from Rachel on word types were less supportive. First, there were missing dorsal consonants, but this did not eliminate the use of back vowels. Further, target words with dorsal consonants showed fronting, e.g. goat [do]. This led to a greater than expected use of coronal consonants with back vowels, a pattern that violates the prediction of FCT. This substitution pattern also indicates the influence of perception on speech production, another result not predicted by FCT.

Concerning homonyms, it was predicted that the preferred syllable shapes should occur greater than expected by chance. In fact, the opposite result was found for Samuel. One possibility would be that the prediction was poorly formulated, and that
FCT predicts that the preferred syllables should occur less with homonyms. While the results would support this revised prediction, it is not obvious that it is anything but ad hoc. If children prefer the syllables ‘di’, ‘ba’, ‘gu’, why would these syllables not form the basis for the child’s homonyms? The one possibility that arises is that homonyms are more likely when children cannot produce unique syllables for a word. That is, non-preferred syllables are more difficult for the child to produce, and therefore are more prone to homonymy. If this proves to be the case, then the homonym data from Samuel can be interpreted as providing evidence for FCT. The data for Rachel did not support this possibility, but the fact that she used few homonyms overall needs to be taken into consideration.

The results for both children on CV phonotactics show high utilization of all possible syllable shapes. For Samuel, 9 of the 30 (30%) possible syllable shapes would have been considered preferred, yet 27 or 90% actually occurred, three times the predicted number. For Rachel, 7 of the 25 possible syllable shapes would have been preferred (28%), yet 22 or 88% were actually found. These high rates of syllable use indicate that the two children were making the most of their limited inventories, despite the articulatory biases toward preferred CV interactions.

Lastly, both children showed evidence of phonological organization and perceptual influence. Samuel showed words that retained the consonant of the first syllable and the vowel of the second syllable, as in Murphy [mi]. This pattern has also been reported in Johnson, Lewis and Hogan (1997), who report a case study of a young child whose early words were highly constrained to a CV syllable. They report numerous examples of his use of sounds from the first and last syllables of multisyllabic targets, e.g. piano [po], candies [kiz]. Despite a high rate of homonymy, Samuel also showed some cases of keeping potential homonyms distinct in the pairs daddy happy and mama grandma by the use of distinctive stress. Samuel also showed an unusual production of [dædɪbʌ] for the dog Kip. Whatever its origin, the form showed phonological creativity not expected if children simply map from target words to preferred syllables. Rachel showed fewer phonological examples, but her fronting pattern of coronals with back vowels and her retention of syllables showed the influence of perception.

As stated in the Introduction, the role of perceptual factors is a critical difference between linguistically-based theories of child phonology versus articulatorily-based ones such as FCT. Bunta, Davidovich and Ingram (2006) provided evidence from the study of a Hungarian-English bilingual girl of the critical role of perception in early phonological acquisition. It was found that the complexity of her word productions significantly varied between the two languages, with the Hungarian words being longer than the English ones. Her proximity to the target words in each language, however, was the same. The authors proposed that the focus over the years on phonological acquisition as a process of increasing articulatory complexity may be misleading, and that phonological acquisition may be more driven by perceptual factors than is sometimes credited.

The results of the present study only provide limited support for FCT. Limited support has also been reported in the small number of studies done by researchers other than the proponents of the theory. Table 10 summarizes some of the studies that have examined FCT. The results provide somewhat of a mixed bag, with no striking trend. They provide weak support for the coronal-front and dorsal-back interactions, with some counterevidence for each. They also support two patterns not
predicted by the FCT, a labial-back interaction, and a coronal-back interaction. We saw the latter in Rachel’s data due to her fronting of velar consonants. Virtually no data support the interaction between labials and central vowels. These contrasting findings can be interpreted as due to different subjects, methods of analysis such as the classification of vowels, and languages. It can also be interpreted that these patterns can be highly influenced by the input language (c.f. Ingram, 2008), and the individual differences found in phonological acquisition.

Table 10. A comparison of five studies on CV interactions

<table>
<thead>
<tr>
<th>Study</th>
<th>Analyses &amp; Results</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vihman (1992)</td>
<td>Comparison of English, French, Japanese, and Swedish children; partial support for labial-central vowel interaction, and for dorsal-back vowels; little support for the coronal-front vowel interaction;</td>
</tr>
<tr>
<td>Tyler &amp; Langsdale (1996)</td>
<td>Comparison of 9 English children; partial support for the dorsal-back interaction; little support for either labial-central vowels or for coronal-front vowels;</td>
</tr>
<tr>
<td>Chen &amp; Kent (2005)</td>
<td>Study of 24 Mandarin-speaking children; support for coronal-front, and velar-back, but not for labial-central; support for a language specific labial-back interaction;</td>
</tr>
<tr>
<td>Stoel-Gammon &amp; Peter (2008)</td>
<td>Analysis of target words in children from three sources; little support for any of the predicted interactions; partial support for an interaction between coronal consonants and back vowels.</td>
</tr>
<tr>
<td>Ingram &amp; Ingram (present study)</td>
<td>Partial support for FCT predicted interactions in word types for one child but not for the other child.</td>
</tr>
</tbody>
</table>

The study in Table 10 that closest approximates the analyses conducted in the present study on word types is Tyler and Langsdale (1996). Tyler and Langsdale studied children at three ages, 1;6, 1;9, and 2;0s, with three children followed longitudinally, and two additional cross-sectional subjects at each age range. They did not calculate ratios of observed to expected CV occurrences, but their frequency tables (Tables 5, 6, 7) allow extrapolation of their data for this purpose. Table 11 presents the results of our analysis of their data, with the FCT preferred syllables in bold.

The chi-squared analyses for 1;6 and 1;9 show no CV interactions occurring outside the expected ranges. A significant difference was found at 2;0, but only the dorsal-back interaction was in the predicted direction. Coronal consonants showed higher than expected occurrences with back vowels, and so did bilabial consonants with front vowels. These results from an analysis of word types versus tokens support the result of the present study that FCT does not predict the distributions found in the lexical types of children’s vocabularies.
Table 11. Frequencies and ratios of observe to/expected occurrences from Tyler and Langsdale (1996) for ages 1;6, 1;9 and 2;0

<table>
<thead>
<tr>
<th>Age (Mo)</th>
<th>Vowels</th>
<th></th>
<th></th>
<th></th>
<th>Vowels</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>18</td>
<td>Coronal</td>
<td>13</td>
<td>13</td>
<td>37</td>
<td>0.87</td>
<td>1.03</td>
<td>1.06</td>
</tr>
<tr>
<td></td>
<td>Labial</td>
<td>32</td>
<td>26</td>
<td>66</td>
<td>1.06</td>
<td>1.05</td>
<td>0.96</td>
</tr>
<tr>
<td></td>
<td>Dorsal</td>
<td>5</td>
<td>2</td>
<td>11</td>
<td>1.14</td>
<td>0.55</td>
<td>1.10</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>$\chi^2$ (4, N=9) = 0.89, p = .93</td>
<td></td>
<td></td>
</tr>
<tr>
<td>21</td>
<td>Coronal</td>
<td>33</td>
<td>19</td>
<td>40</td>
<td>0.98</td>
<td>1.05</td>
<td>0.99</td>
</tr>
<tr>
<td></td>
<td>Labial</td>
<td>72</td>
<td>36</td>
<td>79</td>
<td>1.05</td>
<td>0.98</td>
<td>0.96</td>
</tr>
<tr>
<td></td>
<td>Dorsal</td>
<td>16</td>
<td>10</td>
<td>26</td>
<td>0.84</td>
<td>0.98</td>
<td>1.14</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>$\chi^2$ (4, N=9) = 1.06, p = .86</td>
<td></td>
<td></td>
</tr>
<tr>
<td>24</td>
<td>Coronal</td>
<td>28</td>
<td>22</td>
<td>64</td>
<td>0.80</td>
<td>0.96</td>
<td>1.14</td>
</tr>
<tr>
<td></td>
<td>Labial</td>
<td>81</td>
<td>45</td>
<td>95</td>
<td>1.19</td>
<td>1.02</td>
<td>0.88</td>
</tr>
<tr>
<td></td>
<td>Dorsal</td>
<td>14</td>
<td>13</td>
<td>37</td>
<td>0.71</td>
<td>1.02</td>
<td>1.18</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>$\chi^2$ (4, N=9) = 9.86, p = .05</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

To conclude, FCT is accurate in its findings that children’s early productions when tokens are considered show preferred consonant-vowel interactions. The present study has found that these preferred interactions are carried over to word types to a degree, particularly for Samuel. These preferred interactions, however, were insufficient to capture the broader range of phonological patterns found in the children’s words. In particular, they did not predict the broad range of CV interactions found where each child optimized most of their possibilities within an extremely simple set of possible vowels and consonants, nor the cases where perceptual influence was evident. It is not proposed that data from two children can lead to definitive conclusions, but we propose that such results are likely exemplary. That is, we predict that similar analyses conducted on a comparable data set will provide additional examples of phonological patterns beside those predicted by FCT. A brief review of other studies similarly show mixed results. It is concluded that FCT predicts articulatory tendencies of CV interactions but is too narrow an account of early phonological acquisition. In our view, early phonological acquisition is a combination of phonological organization, the influence of perceptual factors, and the relaxing of the articulatory constraints that have been operative earlier during babbling and explained by FCT.
References
Hills, E. 1914. The speech of a child two years of age. *Dialect Notes* 4, 84-100.
COPING WITH THE EFFECTS OF COARTICULATION: SPANISH AND PORTUGUESE LISTENERS’ PERCEPTION OF NASAL VOWEL HEIGHT

C. Elizabeth Goodin-Mayeda
University of Houston
e-mail: cegoodin@central.uh.edu

Abstract
Because segments in the speech signal are not produced in isolation, but rather in fluid speech, it is inevitable that some degree of coarticulatory overlap occurs between neighboring segments. A common example of this is the overlap in velar aperture that occurs during the articulation of a vowel that is adjacent to a nasal consonant, which results in vowel nasalization. While some research supports that native language experience has little influence on a listener’s ability to compensate for the acoustic consequences of coarticulation (Mann, 1986), there is also evidence that linguistic experience may play a role in the ability to “undo” the effects of coarticulation, specifically in the case of vowel nasalization (Krakow, Beddor, Goldstein, & Fowler, 1988). The current study examines the role of linguistic experience in coping with the effects of coarticulation by testing native Spanish- and Portuguese-speakers’ perception of nasal vowel height (as compared to their oral counterparts) in different contexts: (1) before an adjacent nasal consonant, when the effects of nasalization could be attributed to the nasal consonant, and (2) when no nasal consonant is present. Because Spanish and Portuguese differ in their distribution of nasalized vowels, subjects are expected to resolve the effects of nasal coupling (i.e., the addition of resonant frequencies due to opening the nasal cavity) differently if linguistic experience is a factor in a listener’s ability to compensate for effects of coarticulation. Data support that the ability to perceptually compensate for nasal coupling is dependent on native language experience. Results are discussed in relation to the perception of other coarticulatory effects and diachronic phonological change.

1 Introduction
Despite the tremendous phonetic variability among different pronunciations of the “same” sound, whether due to differences between speakers or differences between tokens spoken by the same speaker, listeners seem able to ignore phonologically non-relevant variations in the speech signal, demonstrating perceptual constancy. Such perceptual constancy in the face of variable stimuli has been noted for other senses as well, particularly sight. For example, objects that are farther away look smaller, yet they are perceived as the same object (Gregory, 1966). For the listener, besides factors such as rate of speech and differences in pitch and amplitude, variability is introduced as a result of the phonetic environment in which a sound is uttered. This variability is due to the fact that individual segments are not produced in isolation, but rather in fluid speech, which inevitably leads to some degree of coarticulation with neighboring sounds. For example, acoustic analyses show that
the noise spectrum of fricative consonants varies with the following vowel due to anticipatory lip rounding of vowels such as [o] and [u], which results in lowering of the fricative noise spectrum (Fujisaki & Kunisaki, 1978; Heinz & Stephens, 1961). In spite of this acoustic variability, Mann and Repp (1980) have noted for English, and Kunisaki and Fujisaki (1977) for Japanese, that listeners are able to perceptually compensate for the effects of anticipatory lip rounding in the relevant phonetic environments. They found that when presented with fricative noise from the [∫]/ch continuum, listeners perceived [s] more often before [u] than [a], indicating that they were adjusting for the coarticulatory effects of rounded vowels on preceding fricatives (Mann & Repp, 1980). Similar findings of perceptual adjustment have been reported for other coarticulatory processes, such as nasal place assimilation (Mitterer & Blomert, 2003) and changes in vowel acoustics between two coronal consonants (Ohala, 1986).

There is evidence that listeners are not always able to perceptually cope with the effects of coarticulation. An example of this is vowel nasalization, which has been shown to affect perceived vowel quality, in particular vowel height, due to the presence of an extra peak in the region associated with vowel height (Beddor, 1993; Chen, 1997; Krakow et al., 1988). According to Chistovich and Lublinskaya (1979), when two adjacent peaks in a vowel spectrum are close in frequency, listeners determine vowel quality based on some weighted average of the two peaks, or center of gravity, and not on the frequency of individual peaks. Thus, an extra nasal peak in the region associated with vowel height is not perceived on its own, but rather concurrently with the oral formant, resulting in a shift in the perceived height of nasal vowels. Wright (1975, 1986) studied American English listeners’ perception of oral and nasal vowels (e.g., [o] vs. [ō]), and found that listeners perceived high vowels and some mid-vowels as lower and low vowels as higher. This follows from the magnetic effects of the nasal formant, which appears at frequencies between the middle to lower extreme of the F1 region of the spectrum. The weighted average of a low frequency oral formant and a slightly higher frequency nasal pole results in listeners perceiving a peak that is slightly higher in frequency than the oral formant, which corresponds to perceiving the vowel quality as slightly lower. Likewise, the “center of gravity” between a high oral formant F1 plus a mid-frequency nasal pole will result in the perception of a slightly lower peak, corresponding to a somewhat higher vowel. According to Ohala (1981), failure to perceptually factor out the spectral effects of nasal coupling (i.e., the nasal resonances introduced when the nasal cavity is open to the oral cavity) has lead to vowel height changes in many languages.

An issue that remains unclear is the role of language experience in compensating for the acoustic effects of coarticulation. Mann (1986) proposes that while there is a language-specific level of speech perception in which speech sounds are represented in accordance with the phonological system of a language to make phonemic distinctions, native language experience has less influence on the ability to compensate for the acoustic consequences of coarticulation. She found that, despite the fact that they were unable to reliably distinguish between English /l/ and /ɾ/, Japanese listeners, like English listeners, were sensitive to coarticulatory effects of /l/ on the /d/-/g/ continuum. English and Japanese listeners were tested on their perception of natural /al/ and /ar/ utterances placed in front of a synthetic continuum ranging from /da/ to /ga/, resulting in the nonce forms /al-da/, /al-ga/, /ar-da/, /ar-ga/.
The results showed that when the preceding utterance ended in /l/, both Japanese and English listeners reported more /ga/ percepts than when the preceding utterance ended in /r/.

In contrast to Mann (1986), Krakow, Beddor, Goldstein and Fowler (1988) assume that the ability to compensate for the effects of nasal coupling depend on native language experience. They found that English speakers were able to “undo” the coarticulatory effects of nasal coupling with the vowel under certain conditions, namely, when the listener could attribute nasalization to an adjacent tautosyllabic nasal consonant. English speakers were tested on their perception of the /el/-/æ/ contrast in oral ([bɛd]–[bæd]), non-contextual nasal ([bɛd]–[bɛd]) and contextual nasal conditions ([bɛnd]–[bɛnd]). Listeners showed a lowering effect (that is, they reported hearing /æ/ more often) for only the non-contextualized nasal vowels, while their perception of contextualized nasal vowels did not differ significantly from that of oral vowels.

Krakow et al. take their findings to suggest that English speakers, due to their lack of experience with phonemic nasal vowels, interpreted the spectral effects of nasalization in the non-contextual nasal vowel condition as a difference in vowel height. In other words, they assume that English speakers’ inability to assess nasal vowel height in the non-contextual condition is a result of their L1 experience (or inexperience). However, based on just these results, it is not clear that coarticulatory influences on the perception of nasal vowels are language specific, as Krakow et al. (1988) suggest. This is because, if coarticulatory influences on the perception of nasal vowels were universal, speakers of all languages would be expected to accurately perceive nasal vowel height only when the effects of nasalization are attributable to an adjacent nasal consonant, i.e., the coarticulatory effects of nasalization would surface in the same contexts as Krakow et al. found for English speakers. Thus, the English speaker results are congruent with both possibilities, that compensation for coarticulation is dependent on native language experience or that it’s universal.

In order to examine the ability of listeners to perceptually undo the effects of nasal coupling, it is necessary to test speakers of languages with different patterns of vowel nasalization. If coarticulatory influences on the perception of nasal vowels are language specific, we would expect cross-linguistic differences in the context effects on the perception of nasal vowels. Specifically, speakers of a language that has phonemic and allophonic nasal vowels (i.e., contextual and noncontextual nasal vowels) would be expected to accurately perceive nasal vowel height in both contextual and noncontextual nasal conditions, due to their experience with nasal vowels in both environments. By the same token, speakers of a language that does not nasalize vowels in any context should always resolve the spectral effects of nasalization in terms of tongue/jaw configuration (i.e., there should be lowering or raising effects for both contextual and noncontextual nasal vowels). Note that the question is not whether listeners perceive the difference between oral and nasal vowels, as this distinction seems to be one that speakers of all languages are able to make (Beddor & Strange, 1982). The question here is whether or not listeners are able to accurately perceive the height of a nasal vowel as compared to its oral counterpart in certain contexts, namely before a tautosyllabic nasal consonant or not.

The purpose of the current study is to examine the role of native language experience in coping with the effects of coarticulation, specifically nasal coupling.
The patterns of nasalization in Spanish and Portuguese present an opportunity to investigate this, as they are different from each other and from those of English.

2 Vowel nasalization in Spanish and Portuguese

Although so-called ‘radical’ dialects of Spanish show phonological nasalization of vowels in prenasal contexts, most ‘conservative’ dialects do not\(^1\). Cross-linguistic studies have shown that lowering the velum before articulation of nasal consonants necessarily involves some amount of overlap with the articulation of the preceding vowel, and that the length of this overlap varies across languages (Solé, 1992). The fact that cross-linguistic variability in degree of vowel nasalization exists suggests that something more than just mechanics of articulation is at play. That is, if anticipatory nasalization were simply due to constraints on speech production, we would not expect languages to differ in the amount of time needed to lower the velum before articulation of a nasal consonant. Solé (1992) has argued that in languages with limited nasalization, vowels are targeted as oral and any overlap of velar opening is a result of physiological time constraints involved in lowering the velum, whereas in languages with extensive nasalization, such as English, nasalization is “intended by the speaker, part of the programming instructions and not a function of physiological constraints of the vocal organs” (Solé, 1992: 30).

Using a nasograph, Solé (1992) showed that, regardless of speech rate, Spanish speakers lower their velum at roughly the same time before the onset of a nasal consonant, about 20ms before the onset. Furthermore, Spanish speakers show very little overlap between articulation of the vowel and lowering the velum for the nasal consonant. In contrast, American English speakers lowered the velum at or before the onset of the prenasal vowel across all speech rates; thus, the voiced portion of the vowel was completely nasalized. These results support that for Spanish speakers, lowering of the velum before the onset of a nasal consonant is due to mechanical demands of articulation, and thus is not a phonological process, but rather a phonetic outcome due to physiological time constraints involved in lowering the velum (Solé, 1992). Because the overlap between articulation of the vowel and velar aperture is minimal (especially compared to languages like English), and because Spanish does not have distinctive nasalization of vowels, it might be expected that Spanish speakers lack the experience (or the necessity) of teasing out the coarticulatory effects of vowel nasalization when followed by a nasal consonant.

Portuguese\(^2\), in contrast, has both allophonic and contrastive nasalization. Allophonic nasalization occurs when stressed vowels appear before a nasal consonant (e.g., [ˈɡõmɐ] goma ‘starch’ vs. [ɡoˈmadu] gomado ‘starched’). The lexical representation of contrastive nasal vowels in Portuguese is the subject of much controversy among Portuguese linguists. The existence of minimal pairs like lâ ‘wool’ and lá ‘there’ have prompted many linguists (Hall, 1943; Hammarström, 1962 and others) to analyze nasalization in this context as part of the lexical representation of the vowel, while others call for analyzing contrastive nasal vowels as syllables closed by some nasal element, either a nasal consonant (Quicoli, 1990; Reed & Leite, 1947) or a nasal mora (Wetzels, 1997). Regardless of whether contrastive nasal vowels are analyzed as oral vowels followed by nasal mora or underlying nasal vowels, on the surface there are minimal pairs such as [sej] (sei ‘1 sg. know’) vs. [sẽj] (sem ‘without’) and [u] livro (o livro ‘the book’) vs. [ũ] livro (um livro ‘a book’); thus, Portuguese speakers have experience perceiving the oral-
nasal vowel contrast even when nasal vowels are not followed by a nasal consonant in production.

Returning to the research question, we can make specific predictions regarding Spanish-speakers’ and Portuguese-speakers’ perception of nasal vowel height. If the effects of coarticulation on nasal vowel perception are universal, crosslinguistic effects of context should mirror the results found by Krakow et al. (1988) for English speakers. Specifically, here speakers of both Spanish and Portuguese would be expected to “tease out” the coarticulatory acoustic effects of nasalization only when there is an adjacent nasal consonant to which these effects can be attributed (i.e. in the contextual nasal condition). If, however, the ability to compensate for the effects of nasal coupling is language-dependent, Spanish speakers should perceive a shift in nasal vowel height, as compared to oral vowel height, regardless of context, while Portuguese speakers, in contrast, should accurately perceive nasal vowel height in all contexts, since they have experience with both contextual and noncontextual nasal vowels. Figure 1 below illustrates the predictions of these two hypotheses. Note that the predictions of both hypotheses are the same for English speakers, and that these predictions are consistent with Krakow et al.’s (1988) findings.

<table>
<thead>
<tr>
<th>Universal</th>
<th>Language-dependent</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Accurately perceive contextual nasal vowel height</strong></td>
<td><strong>Accurately perceive noncontextual nasal vowel height</strong></td>
</tr>
<tr>
<td>English</td>
<td>✓</td>
</tr>
<tr>
<td>Spanish</td>
<td>✓</td>
</tr>
<tr>
<td>Portuguese</td>
<td>✓</td>
</tr>
</tbody>
</table>

*Figure 1. Comparing predictions of hypotheses that coarticulatory effects on perception of nasal vowel height are universal vs. language dependent*

3 Methodology
3.1 Subjects

Twelve speakers of Castilian Spanish and twelve speakers of Brazilian Portuguese were tested. All subjects were native speakers of their prospective language and reported no hearing impairments. An effort was made to ensure that the subject groups were as homogenous as possible in terms of their linguistic background, both first (L1) and second language (L2). The Spanish subjects were all speakers from Barcelona, Spain. While it may seem strange to choose to conduct an experiment of this sort in an area of bilingualism, it was important to ensure that the assumptions regarding speech production were true, in order to test perception of the same phenomena. In that respect, what better place to test perception of nasalized vowels than in the very region in which Solé collected her production data? Although it is not expressly mentioned in her study, Solé’s subjects, as the subjects of this study, were most likely bilingual or at least proficient speakers of Catalan. Although
Catalan phonology is very different from that of Spanish, experience with Catalan should pose no problems for studies involving nasalization of vowels since, at least in this respect, Catalan and Spanish are similar (Wheeler, 2005). In addition, all subjects of the current study reported being Spanish dominant.

It was somewhat more difficult to maintain such a high level of homogeneity within the Portuguese group. All subjects were Brazilians living in Los Angeles and had learned English as adults. All reported using Portuguese on a regular basis. Subjects were from Porto Alegre (n=2), São Paulo (n=3), Rio de Janeiro (n=3), Minas Gerais (n=2), Recife (n=1) and Maceió (n=1).

3.2 Stimuli

Because the subjects of this study are from different language backgrounds, non-words that conform to the phonotactic constraints of Spanish and Portuguese were chosen to test the /u/-/o/ distinction. The stimuli for this experiment consisted of synthesized non-words in which vowel height and degree of nasalization were manipulated. The choice of synthetic or modified natural speech tokens (or hybrid tokens) depends largely on the type of distinction being examined, and each comes with its own advantages and disadvantages. Unfortunately, natural speech tokens, unlike synthetic tokens, do not permit total control over variations in the acoustic signal. While synthetic stimuli may not succeed in representing some perceptually relevant properties of the signal, a positive correlation between perception of natural and synthetic speech has been attested in some studies (Werker & Lalond, 1988; Yamada & Tohkura, 1992).

Stimuli were synthesized using the program Synthworks, a KLATT based parallel speech synthesizer with up to 48 manipulable parameters. Natural tokens of the endpoints [gos], [gus], [gõns] and [gũns] provided the basis for synthesizing the basic shape of the stimuli, including parameters such as segment length, formant transitions for initial /g/, voicing cues, total utterance length, etc. The fundamental frequency (F0) of all utterances began at 100 Hz and fell linearly to 85 Hz over the first 200 ms of the utterance and remained there until the end of the utterance. The vowels were generated by first synthesizing the endpoints /u/ and /o/ and then manipulating the first formant (F1) in equal increments to synthesize five intermediate shapes in order to ultimately create a seven-step vowel continuum from /u/ to /o/ (see appendix A for formant frequencies of stimuli). The seven vowel heights were used in three synthetic continua: oral [gos]-[gus], contextual nasal [gõns]-[gũns] and noncontextual nasal [gõs]-[gũs]. While all of these words are nonce forms in Spanish and Portuguese, the tokens [gus] and [gũns] are phonetically similar to English ‘goose’ and ‘goons,’ respectively, which could be potentially problematic in the case of the Portuguese speakers tested here, since they are L2 learners of English. I will argue in the discussion that these tokens should not pose a problem for the Portuguese speakers, because, among other reasons, even the endpoint vowel /u/ is acoustically quite different from a typical American English /u/, as it was synthesized as rounded and monophthongal.

For the two nasal conditions, five degrees of nasalization were synthesized in order to account for possible speaker variability in the degree to which vowels must be nasalized to be perceived as nasal and also to permit examination of this effect on perceived vowel height. The primary cues for nasal coupling are reduction in amplitude of the first formant, an increase in F1 bandwidth of up to 60 Hz as well as the introduction of two nasal peaks, one at around 1000 Hz and another between 250
and 450 Hz (Chen, 1997). Nasalization was synthesized by introducing nasal poles and zeros at varying frequencies, depending on the level of nasalization. In addition, the amplitude of the first oral formant was decreased in the nasal continuum. Nasality judgments have been correlated with the magnitude of $F1$ amplitude reduction (Chen, 1997; House & Stevens, 1956), thus, the lower and middle degrees of nasalization were synthesized with a 6 dB, 7 dB and 8 dB reduction in $F1$ amplitude, respectively, and the highest two degrees of nasalization were synthesized with a reduction in $F1$ amplitude of 8 dB. Finally, the amplitude of the nasal pole increased from 30 dB to 60 dB over the duration of the (nasal) vowel. (See Appendix.)

The stimuli consisted of 77 utterances in all, separated into 3 sets as shown in Table 1. After all stimuli were synthesized, they were converted into .wav audio file format.

### Table 1. Three stimulus sets synthesized for identification task.

<table>
<thead>
<tr>
<th>Stimulus sets</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Oral condition [gus]–[gos]</td>
</tr>
<tr>
<td>7 tokens (seven vowel heights)</td>
</tr>
<tr>
<td>2. Contextual nasal condition [güns]–[gōns]</td>
</tr>
<tr>
<td>35 tokens (seven vowel heights X five degrees of nasal coupling)</td>
</tr>
<tr>
<td>3. Noncontextual nasal condition [gūs]–[gōs]</td>
</tr>
<tr>
<td>35 tokens (seven vowel heights X five degrees of nasal coupling)</td>
</tr>
</tbody>
</table>

#### 3.3 Procedure

Subjects were tested using a forced-choice identification task administered via Praat. The stimuli were counterbalanced across ten blocks so that the subjects heard each stimulus a total of ten times during the experiment. In addition, the order of the stimulus sets was counterbalanced across subjects. Subjects heard the stimuli over headphones, and used the mouse to click on “buttons” on the computer screen to indicate whether the vowel they had heard was [u] or [o]. To avoid the possible confound of orthography, the entire word was not spelled out. These response choices posed no difficulty for the subjects, because both Spanish and Portuguese orthography are more phonetically reliable than languages like English, for example, and because [u] and [o] correspond to orthographic ‘u’ and ‘o’ respectively in both languages. Each condition was preceded by a short practice session. The stimuli for the practice sessions contained the endpoint vowel heights as well as the next closest vowel height for each set (i.e., vowel heights 1, 2, 6, and 7 from the seven-step vowel continuum). No feedback regarding the subjects’ performance was given after the practice sessions.

Subjects were tested in groups of 1-4 in a quiet environment. The experimenter read instructions in the subjects’ native language, which indicated that they would hear a series of “invented” words, and to imagine that these words were new words of their language. They were instructed to use a mouse to click on the ‘u’ button or ‘o’ button displayed on the computer screen to indicate which vowel they thought best fit the vowel they heard. The pace of the experiment was determined by how quickly the subject chose each response, but the entire experiment generally took between 50 and 70 minutes.
4 Results

Statistical analyses were carried out based on values obtained by a Probit Regression Analysis for the 50% crossover from /u/ to /o/ calculated for each subject’s judgments in each continuum. Crossover values were based on a scale of 1 to 7 corresponding to the seven vowel heights. On this scale, 1 corresponds to the /u/ endpoint (F1 frequency of 225 Hz) and 7 corresponds to the /o/ endpoint (F1 frequency of 450). The 50% crossover values corresponding to Figures 2 and 3 are shown below in Table 2.

Table 2. Mean crossover values for Spanish and Portuguese listeners corresponding to the identification results of the [gus]-[gos], [güns]-[gôns] and [güs]-[gös] continua.

<table>
<thead>
<tr>
<th>Crossover values across language groups</th>
<th>Condition</th>
<th>Spanish Group</th>
<th>Portuguese Group</th>
</tr>
</thead>
<tbody>
<tr>
<td>[gus]</td>
<td>-</td>
<td>gos</td>
<td>3.90</td>
</tr>
<tr>
<td>[güns]</td>
<td>-</td>
<td>gôns</td>
<td>4.71</td>
</tr>
<tr>
<td>[güs]</td>
<td>-</td>
<td>gös</td>
<td>4.53</td>
</tr>
</tbody>
</table>

A two-tailed Analysis of Variance (ANOVA) supports that Spanish speakers reported significantly more /u/ responses in both nasal conditions as compared to the oral condition ($p = .017$), which translates into a raising effect for the nasal vowel conditions only. Post hoc paired-samples $t$-tests confirmed that both contextual and noncontextual nasal vowels were perceived as significantly different (higher) than oral vowels of the same height ($t$ -2.564, $p = .026$, $t$ -4.232, $p = .001$, respectively), but not significantly different from each other ($t$ 1.038, $p = .321$). This corresponds to a perceptual raising effect such that Spanish speakers interpreted the coarticulatory effects of nasalization as a difference in vowel height. Thus, Spanish speakers were unable to perceptually factor out the effects of nasalization in either nasal condition, regardless of whether there was an adjacent nasal consonant to which nasalization could be attributed.

Figure 2 illustrates Spanish speakers’ rate of /u/ responses in each condition. Here the percentage of /u/ responses as a function of vowel height collapsed over degree of nasalization (for nasal stimuli) is shown. The figure represents the pooled responses of all 12 subjects. The figure illustrates that Spanish speakers recorded more /u/ responses in both of the nasal conditions as compared to the oral condition. That is, given two vowels with the same $F1$ frequency, the nasal vowel was more often perceived as /u/ (regardless of its context) as compared to its oral counterpart.
Figure 2 Identification responses to the oral [gVs], contextual nasal [gṼns], and noncontextual nasal [gṼs] continua for Spanish speakers.

Figure 3 shows identification results for all five degrees of nasal coupling (A-E where A is the lowest degree of coupling and E is the highest degree) compared to oral judgments. Here the results of the oral continuum have been replicated in each graph for purposes of comparison. As can be seen, not only did Spanish speakers perceive nasal vowels as higher in both nasal conditions, but they did so across all degrees of nasalization (p = 0.021, 0.017, 0.001, 0.033, and 0.047 for degrees A-E, respectively).

In contrast to the Spanish results, a two-way ANOVA revealed that the number of /u/ responses for the Portuguese speakers was not significantly different in any context (p = .520). Post hoc paired-samples t-tests confirmed that Portuguese speakers’ perception of vowels in the contextual and noncontextual conditions did not significantly differ from those in the oral condition (t = .562, p = .586, t = .966, p = .355, respectively) nor from each other (t = .742, p = .474). These results support that in both nasal contexts, Portuguese speakers were able to perceptually “undo” the effects of nasalization in order to accurately assess the height of the vowel. Figure 3 shows the rate of /u/ responses across vowel height frequency values in each condition for Portuguese speakers. Like Figure 2, Figure 4 represents the pooled responses of all 12 subjects.
Figure 3. Identification functions for contextual [g\~{n}s] and noncontextual [g\~{V}s] nasal continua over five degrees of nasal coupling from slight (level A) to heavy (level E). The identification results for the oral [gVs] condition have been reproduced in each graph for purposes of comparison.
Figure 4. Identification responses to the oral [gVs], contextual nasal [gVns], and noncontextual nasal [gVs] continua for Portuguese speakers.

5 Discussion

The results of the experiment showed that Spanish speakers produced more /u/ responses in both nasal conditions than in the oral condition. Portuguese speaker responses, on the other hand, did not differ significantly in any of the conditions. The fact that Spanish speakers resolved the effects of nasalization as a difference in vowel height supports that they were unable to perceptually undo the effects of nasal coupling in either nasal condition. Portuguese speakers demonstrated that they were able to factor out the effects of nasal coupling with the vowel as evidenced by the fact that they accurately perceived the height of nasal vowels in both nasal contexts. These results in conjunction with the results of Krakow et al. (1988) support that linguistic experience plays a significant role in how speakers resolve the acoustic effects of nasal coupling with the vowel. In each case, listeners were able to perceptually undo the coarticulatory effects only in the environments in which they have experience with nasal vowels.

A key issue to address is why the results of this study support that coarticulatory influences on the perception of nasal vowel height are language-dependent, but coarticulatory influences related to other distinctions seem to be universal (Mann, 1986). Recall that Mann (1986) showed that Japanese speakers were still sensitive to coarticulatory effects of /l/ and /r/ on the /d/-/g/ distinction, in spite of their inability to distinguish between stimuli containing /l/ and /r/.

There are several possible reasons for the incongruity between Mann’s conclusions and those reported here. First, it is possible that some effects of coarticulation are simply different than others, and thus affect perception differently. This may seem like an ad hoc explanation, but let us consider the nature of nasal coupling. While coarticulatory influences on consonants tend to be sequential, the
coarticulatory effects of nasalization are realized simultaneously during the articulation of the vowel. For example, the coarticulatory effects of /l/ on adjacent stop consonant place of articulation consist of neighboring formant transitions of /l/ affecting the perception of formant transitions related to the stop. Nasalization, on the other hand, involves, among other acoustic consequences, the addition of nasal poles and zeros simultaneously during the articulation of the vowel. In this case, the listener’s task may be more difficult, as the spectral consequences of nasalization may be less salient as compared to changes in formant transitions at the segment edge.

It is also possible that Japanese speakers’ experience with other, similar sounds in their language aided them in untangling the effects of coarticulation of /l/. Both Japanese and English speakers demonstrated a similar shift in perception of the /dal/-/gal/ continuum when preceded by /s/ as opposed to /ʃ/ (Mann 1986). Both /l/ and /s/ are similar in that they are produced with the tongue in a relatively forward position, while /r/ and /ʃ/ are produced with a more retracted tongue position. It is possible that Japanese speakers’ sensitivity to the coarticulatory effects of /l/ are due to their experience with the coarticulatory effects of /s/. In this way, they are not factoring out entirely novel acoustic effects of coarticulation.

The effects of nasalization, in contrast, are unlikely to be similar to coarticulatory effects of other sounds. For this reason, Spanish speakers could not make use of knowledge of similar coarticulatory influences of other sounds to compensate for the acoustic effects of nasal coupling in either nasal condition. The same is true for English speakers in the noncontextual nasal condition found by Krakow et al. (1988). Because the coarticulatory effects under consideration in the current study are acoustic effects associated only with nasalization (and not coarticulation with other segments), the results of this study provide stronger evidence that a speaker’s language experience determines the types of coarticulatory influences he or she is able to perceptually factor out.

The role of language experience in perceiving effects of coarticulation has implications in the study of sound change. Ohala (1993) notes the parallels between phonetic variation and sound change. For example, allophonic nasalization of vowels before nasal consonants is the environment which most often results in distinctively nasal vowels via sound change (Ohala, 1993). In this process, an essentially oral vowel precedes a nasal consonant and eventually becomes nasalized. Then the nasal consonant that conditioned the vowel nasalization becomes weakened and is deleted (or absorbed). Sound change occurs when listeners “misperceive” the intended pronunciation of the speaker and this “misperception” leads to a change of norms (Ohala, 1981, 1986, 1990, 1993, 1996). Ohala notes that sound changes that are the result of coarticulatory influences, be they assimilations or other effects resulting from speech production, often occur with the simultaneous loss of the conditioning environment. A change of this nature has already occurred in Brazilian Portuguese. Unlike European Portuguese, Brazilian Portuguese does not have [+low] nasal vowels. Low vowels /e/, /a/ and /o/ raise to mid [ɐ], [ɐ] and [ɒ], respectively, when they become nasal. An example of this is tempo ‘time’, which is pronounced with [ɐ] in Brazilian Portuguese but [ɐ] in European Portuguese. Presumably, the loss/absorption of the nasal consonant that conditioned the nasalization of /e/ resulted in the reanalysis of the height of the vowel in Brazilian Portuguese.
Interestingly, in European Portuguese, which has retained the pronunciation of the nasal consonant, the vowel is still low.

One limitation of the current study is that the Portuguese speakers are also L2 learners of English. This does not pose a problem for the principal findings, that is, that Portuguese speakers perceived nasal vowel height accurately in contextual and noncontextual conditions, since their L2 experience would neither help nor hurt them in terms of perceiving noncontextual nasals; however, there is the question of whether some test tokens that are phonetically similar to English words, specifically \[gus\] and \[gūns\], might have influenced the Portuguese results. Although it is possible that these forms were lexicalized by the Portuguese speakers, there are a couple of reasons why it is unlikely that they “tapped into” these English forms. First, the instructions stated that the test items were not words of any language and directed the subjects to think of the nonce forms as new words of their language. Furthermore, unlike English \(/u/\), which is unrounded, usually diphthongized and starts as a more central vowel, the formant frequencies that were synthesized for \(/u/\) were monophthongal and corresponded to a high back rounded vowel. In fact, an attempt to collect English speaker data was made, but pilot data showed that English speakers had a remarkably difficult time labeling even the endpoints \([u]\) and \([o]\) (also synthesized as a monophthong). It is unclear whether this was due to the fact that English \(/u/\) and \(/o/\) are acoustically different than the synthesized vowel endpoints or to difficulties with labeling these sounds as ‘u’ and ‘o’, given the fact that English orthography is less transparent than that of Spanish or Portuguese. An additional challenge of including English speakers was creating stimuli that conformed to phonotactic/phonological constraints of all three languages and yet were not actually words. In spite of these obstacles, it would be interesting to test speakers of all three languages, Spanish, Portuguese and English, using the same stimuli. This, unfortunately, will be left for future research.

6 Conclusion

The current study sought to investigate the role of native language experience in perceptually compensating for the acoustic effects of vowel nasalization in different contexts. The data demonstrate that while Spanish speakers were unable to perceptually factor out the effects of nasalization in either contextual or noncontextual conditions, Portuguese speakers were able to do so in both conditions. These data in conjunction with the English speaker results from Krakow et al. (1988) support that native language experience does indeed play a role in the ability to perceptually undo the acoustic effects of coarticulation, at least in the case of nasal coupling. This has important implications for the study of diachronic sound change as well as models of speech perception and L2 acquisition of phonology. For example, it stands to reason that failure to factor out the effects of coarticulation due to lack of experience with such effects could result in additional L2 learner difficulty in perceiving L2 distinctions in certain contexts. Future research might explore how the results of this study relate to Second Language Acquisition (SLA), as well as examine the role of linguistic experience in perceptually compensating for other types of coarticulation.
References


### Appendix

Table 1 shows the frequency values for the first three oral formants in all 7 stimuli.

<table>
<thead>
<tr>
<th>Stimulus</th>
<th>F1 (Hz)</th>
<th>F2 (Hz)</th>
<th>F3 (Hz)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>225</td>
<td>750</td>
<td>2800</td>
</tr>
<tr>
<td>2</td>
<td>262</td>
<td>750</td>
<td>2800</td>
</tr>
<tr>
<td>3</td>
<td>300</td>
<td>750</td>
<td>2800</td>
</tr>
<tr>
<td>4</td>
<td>337</td>
<td>750</td>
<td>2800</td>
</tr>
<tr>
<td>5</td>
<td>375</td>
<td>750</td>
<td>2800</td>
</tr>
<tr>
<td>6</td>
<td>412</td>
<td>750</td>
<td>2800</td>
</tr>
<tr>
<td>7</td>
<td>450</td>
<td>750</td>
<td>2800</td>
</tr>
</tbody>
</table>
Table 2 shows the frequency of the nasal pole (fnp) and zero (fnz) as well as the
amount of reduction in F1 amplitude (AF1). In addition, the amplitude of the nasal
pole increased from 30 dB to 60 dB over the duration of the (nasal) vowel.

Table 2. Frequency of nasal pole zero pairs and amount of amplitude reduction of
oral F1 for all degrees of nasalization.

<table>
<thead>
<tr>
<th>Degree of Nasalization</th>
<th>fnp (Hz)</th>
<th>fnz (Hz)</th>
<th>AF1 (dB)</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>250</td>
<td>310</td>
<td>-5</td>
</tr>
<tr>
<td>b</td>
<td>265</td>
<td>365</td>
<td>-6</td>
</tr>
<tr>
<td>c</td>
<td>290</td>
<td>420</td>
<td>-7</td>
</tr>
<tr>
<td>d</td>
<td>305</td>
<td>475</td>
<td>-8</td>
</tr>
<tr>
<td>e</td>
<td>315</td>
<td>528</td>
<td>-8</td>
</tr>
</tbody>
</table>

Notes
1. Note that some dialects of Spanish, e.g., Caribbean dialects, exhibit nasalization
patterns more like those found in English; here we are only concerned with patterns
Also, vowels in both conservative and radical dialects tend to become nasalized
when between nasal consonants, as in comūn, where the /u/ is nasalized (Schwegler
et al., 2010).
2. Unless otherwise stated, all references to Portuguese refer to Brazilian
Portuguese.
3. It should be noted that while Spanish allows syllables of the shape CVNs (e.g.,
cons-tan-te), there are very few, if any, monosyllabic words with this shape. A
possible exception is the pronunciation [tons] for ‘entonces.’
4. As one reviewer points out, this may have influenced the results in the sense
that subjects were lead to rely on their native language phonology, instead of using a
‘naïve’ ear. The motivation for this phrase in the instructions was to rule out the
possibility that subjects thought of the words as English words or as nonlinguistic
gibberish.
5. Here I follow Quicoli (1990) and others, who analyze /e/, /a/ and /ʌ/ as
phonologically low vowels in Portuguese.
THE PRODUCTION OF POST-ALVEOLAR VOICELESS FRICATIVES AND AFFRICATES BY BILINGUAL SPANISH- AND ENGLISH-SPEAKING CHILDREN AND THEIR MONOLINGUAL PEERS: A PILOT STUDY

Ferenc Bunta¹, Christina DiLuca¹, and Lee Branum-Martin²
¹Department of Communication Sciences and Disorders, The University of Houston
²Texas Institute for Measurement, Evaluation, and Statistics, The University of Houston
e-mail: fbunta@Central.UH.EDU

Abstract

Purpose: The present study investigated the productions of the voiceless post-alveolar fricative and affricate by bilingual English- and Spanish-speaking, monolingual English-speaking, and monolingual Spanish-speaking children.

Method: Fourteen typically developing 3- to 6-year-old children participated in this study: 5 monolingual English-speaking, 5 bilingual English- and Spanish-speaking, and 4 monolingual Spanish-speaking children. Single word speech samples targeting word-initial voiceless post-alveolar affricates and fricatives were obtained to investigate the phonemes and their distinguishing acoustic cues (i.e., durations and rise times).

Results: Bilingual children displayed more examples of affrication of the fricative than their monolingual English-speaking peers. Only one statistically significant difference was found differentiating bilingual and monolingual children’s /ʃ/ in the word “shovel”. The other contrasts tended to have medium to large effect sizes but were not statistically significant.

Conclusions: Effect sizes seemed to suggest that monolingual English-speaking children tended to have longer voiceless post-alveolar fricatives than their bilingual peers speaking English. In general, monolingual English-speaking children also had longer fricative portions in their affricates than their bilingual peers, who, in turn tended to have longer fricatives in the affricates than their monolingual Spanish-speaking peers. Future investigations should further expand and explore how voiceless post-alveolar affricates and fricatives are acquired by bilingual children and their monolingual peers so as to gain a more complete and accurate picture of phonological development in bilingual and monolingual children.

1 Introduction

The acquisition of fricatives and affricates poses a special challenge to bilingual Spanish- and English-speaking children, because these sounds are not among the early acquired phonemes in either English (Ingram, Christensen, Veach, & Webster, 1980; Shriberg, 1993) or Spanish (Fabiano-Smith & Goldstein, 2010a), and the two languages also differ significantly in their fricative and affricate inventories and how those phonemes are used in each language (Goldstein, 1995). For example, while
both Spanish and English have their versions of the voiceless post-alveolar affricate, \(/\text{ʧ}/\), most dialects of Spanish lack the post-alveolar fricative (cf. Goldstein, 1995). Thus, the linguistically marked nature of post-alveolar fricatives and affricates coupled with cross-linguistic differences make their acquisition difficult for bilingual Spanish- and English-speaking children. The present study focuses on the production of \(/\text{ʧ}/\) in English and Spanish and \(/\text{ʃ}/\) in English by four- to six-year-old bilingual children and their monolingual peers, using both phonetic transcriptions and acoustic analyses.

1.1 Properties of Post-Alveolar Fricatives and Affricates in English and Spanish

The phonologies of American English and Mexican Spanish differ in important ways regarding their fricatives and affricates, but they also display some similarities. In terms of phoneme inventories, English has a more extensive set of consonant phonemes than Spanish (cf. Goldstein, 1995). There are 24 consonants in American English (Smit, Hand, Freilinger, Bernthal, and Bird, 1990; Sander, 1972), compared to 18 in “standard” Spanish (Bedore, 1999; Goldstein, 1995). Nonetheless, there are 14 consonants that occur in both languages (/p d k b d g f s ʧ j l m n/) so the majority of the consonants have analogs in Spanish and English (Goldstein, 1995).

Considering fricatives and affricates, English has 9 that belong to the former category, /h f v s z ʃ ʒ θ ð/, and two that belong to the latter, /ʤ ʧ/ (Smit et al. 1990; Sander, 1972). Spanish, on the other hand, has only three fricatives, /f x s/, and one affricate, /ʧ/ (Hammond, 2001), so the English system of fricatives and affricates is far more complex than its Spanish counterpart, which has potential ramifications for phonological acquisition in bilingual Spanish- and English-speaking children. Moreover, even the fricatives and the affricate that have analogs in English and Spanish behave differently in the two languages, usually having a more restricted pattern in Spanish. For instance, /ʧ/ and /ʃ/ may occur in word-initial, medial, or final position in English, but they may only occur in word-initial or medial position in Spanish. In English, /ʧ/ is heavily aspirated and is produced with a large amount of friction when used as the onset of a syllable, while the Spanish /ʧ/ is less aspirated and uses relatively shorter friction in a syllable onset (Stockwell & Bowen, 1970).

The voiceless post-alveolar fricative, /ʃ/, is a phoneme in English, but it typically does not occur as a phoneme in most varieties of Spanish. However, some dialects of Spanish do contain [ʃ] typically as a non-contrastive allophone of /ʧ/, such as in parts of Andalucía (Spain), Chihuahua and Sonora (northern Mexico), Panama, and parts of Chile (Hualde, Olarrea & Escobar, 2007). It should also be noted that although this allophonic variation does occur, it is not used frequently nor is it as systematic as the usage of spirantization (Gildersleeve-Neumann, Kester, Davis, & Peña, 2008). Thus, Spanish- and English-speaking bilingual children encounter /ʃ/ in English as a phoneme, but in Spanish, they may only experience /ʃ/ as an allophonic variant of the homorganic affricate, depending on the variety of Spanish they speak.

1.2 Acoustic Characteristics of Post-Alveolar Fricatives and Affricates

Studies using acoustic analyses have found a number of cues that define post-alveolar fricatives and affricates in a variety of languages. Investigating the production of English obstruents, Gerstman (1957) found that in absolute initial position, the release burst associated with affricates, the rise time of fricatives and affricates, and the duration of the fricative noise best differentiate fricatives from affricates. All other things being equal, post-alveolar affricates tend to have a gap,
followed by a burst and then a quickly rising fricative that has about half the rise time and the duration of a regular post-alveolar fricative (Gerstman, 1957). In word-final position, in addition to the burst, rise time and friction duration, Dorman, Isenberg, and Raphael (1980) found that the presence and the duration of the gap, as well as the duration of the preceding vowel, are important cues that differentiate English affricates from fricatives. Dorman et al. (1980) also point out that none of the cues identified by them can be ignored as acoustic curiosities, because each contributes to identifying fricatives and affricates accurately. However, not all studies found all acoustic cues to differentiating affricates from fricatives equally relevant. While largely agreeing with Dorman et al. (1980), Howell and Rosen (1983) found that durational differences of the fricatives and affricates varied depending on the position in the word in which the phoneme occurred. They added that rise time was a salient perceptual cue for the voiceless affricate and fricative set, but not for their voiced counterparts. Furthermore, Kluender and Walsh (1992) found that rise time alone is not a sufficiently reliable cue to distinguishing fricatives from affricates, but the duration of the fricative and the fricative portion of the affricate were more reliable in differentiating the post-alveolar fricative from the affricate. Interestingly, friction duration is a robust enough cue as to not be significantly affected by speaking rate whereas other aspects, such as the duration of the gap and vowels, are more susceptible to speaking rate changes (Maddieson, 1980).

While the majority of the studies on fricative-affricate differentiation focused on English, some studies also included other languages. Maddieson (1980) investigated the properties of affricates and fricatives in word-medial position in Italian, English, and Spanish and found that friction duration was the best measure based on which these languages could be differentiated. The duration of the fricative portion in an English affricate significantly exceeded the friction duration of Spanish post-alveolar affricates; thus this feature warrants exploration in the speech production patterns of bilingual Spanish- and English-speaking children and those of their monolingual peers.

1.3 The Acquisition of Fricatives and Affricates

Numerous studies investigating various languages found that fricatives and affricates are usually not among the early developing phonemes (e.g., Altvater-Mackensen & Fikkert, 2010; Fabiano-Smith & Goldstein, 2010a; Ingram et al. 1980; Kim & Stoel-Gammon, 2011; Shriberg, 1993). Typically, stops are acquired before fricatives and affricates in English (Ingram et al. 1980), Spanish (Fabiano-Smith & Goldstein, 2010a), Korean (Kim & Stoel-Gammon, 2011), and Dutch (Altvater-Mackensen & Fikkert, 2010). Thus, a variety of languages from different language families seem to support a universal pattern of stops occurring before fricatives and affricates. There are also some common general substitution patterns for fricatives and affricates in different languages. For example, stops are among the most common early substitutes for fricatives or affricates in English (Ingram et al. 1980), Dutch (Altvater-Mackensen & Fikkert, 2010), Korean (Kim & Stoel-Gammon, 2011), and Spanish (Anderson & Smith, 1987; Linares, 1981).

While fricatives and affricates tend to be among the middle and later acquired phonemes and some commonalities in the patterns of acquisition in different languages do exist, the exact order of acquisition, substitution patterns, and perceptual attunement are far from uniform across languages. As Kim and Stoel-Gammon (2011) note, children display both universal and language-specific patterns
in their acquisition of obstruents. In fact, Ingram (1988) found that word-initial /v/ was acquired late in English but early in Swedish, Estonian, and Bulgarian, because the function of a sound in a given language affects order of acquisition in important ways. Another example of language-specific acquisition pattern can be found in the fricative substitutions of Japanese-speaking versus English-speaking children. Li, Munson, Edwards, Yoneyama, and Hall (2011) found that Japanese children substitute [ʃ] for /s/ and acquire /ʃ/ much earlier than their English-speaking peers, which may be due to the fact that the acoustic space for the Japanese and English /s/ and /ʃ/ phonemes differ.

It seems that, in general, fricatives and affricates are among the later acquired phonemes in various languages, but the order of their acquisition is moderated by the demands of the specific language being acquired. Thus, it is important to review, specifically, how monolingual and bilingual English- and Spanish-speaking children acquire their fricatives and affricates.

1.3.1 The Acquisition of Fricatives and Affricates by Monolingual English-Speaking Children

As we noted previously, fricatives and affricates are not among the early acquired phonemes in English (Ingram et al. 1980; Sander, 1972; Shriberg, 1993). In categorizing consonants as early-, middle-, or late-acquired, Shriberg (1993) found that /f v tʃ dʒ/ were middle-developing phonemes while /ʃ s z θ/ were fricatives in the late-developing group (/z/ was excluded from the study). Ingram et al. (1980) investigated the acquisition of word-initial /f/, /v/, /tʃ/, /dʒ/, /ʃ/, /θ/, and /z/ by monolingual English-speaking children between the ages of 1;10.22 and 5;11.18. The authors found that among the fricatives studied, /θ/ was acquired the earliest, followed by /ʃ/, /tʃ/, and /dʒ/, and finally, /s/, /z/, and /z/. Consistent substitution patterns included: /ʃ/ -> [s]; /tʃ/ -> [s], [p], or [b]; /v/ -> [b]; /θ/ -> [f]; /s/ -> [θ]; /z/ -> [ð] or [s], /tʃ/ -> [t], [s], or [ts], and /dʒ/ -> [d]. Large individual variation across the children was also found, as well as word familiarity and phonological word complexity effects. Ingram et al. (1980) also concluded that acquisition of the English fricatives and affricates was gradual.

Studies that involve acoustic analyses of fricatives and/or affricates to investigate their acquisition find that even children who are 7 years old do not have completely adult-like representations both in perception and production (e.g., Nittrouer, 1995; Nittrouer & Lowenstein, 2010). Nittrouer (2002; 2006) claims that early in acquisition, children place more emphasis on general vocal tract movements and the resulting changes in the spectral domain, and only later does attention shift to more static cues. Nittrouer, Lowenstein, and Packer (2009) found that children as old as 7 years place more emphasis on global spectral cues than adults, suggesting incomplete acquisition of fricatives. This non-adult-like perceptual weighting of fricative cues also results in poorer perception in 3- to 7-year-old children than in adults (Nittrouer & Lowenstein, 2010). Nissen and Fox (2005) also found that children’s productions of fricatives were not completely adult-like, but there was a significant shift towards more adult-like fricative productions around 5 years of age. For example, /θ/ and /θl/ were longer in children’s productions as compared to adults’ speech patterns; however, /s/ and /ʃ/ tended to be shorter and less well-differentiated in children’s speech samples as compared to adults’ productions. Nissen and Fox (2005) explain these child versus adult production differences with children having still developing vocal tracts and changing (or incomplete) phonological systems.
In comparing studies that use analyses based on phonetic transcription (e.g., Ingram et al. 1980; Shriberg, 1993) to studies that use acoustic analyses (e.g., Nissen & Fox, 2005; Nittouer, 1995), it appears that the latter identify a much later age of fricative acquisition than the former. This discrepancy may be due to the fact that, as Edwards and Munson (in press) note, child forms that are not yet completely adult-like but are often within the adult perceptual category. Studies also differ somewhat regarding the order of fricative and affricate acquisition. For example, Sander (1972) found that /s/ was acquired earlier than /ʃ/, while Ingram et al. (1980) found the opposite, and Shriberg (1993) listed both /s/ and /ʃ/ among the late-developing consonant category. These divergent results may be due to methodological differences (e.g., different acquisition criteria or the use of different test words). In any event, the overall fricative and affricate development patterns seem consistent across the different studies even if specific details differ.

It is clear from the research on English fricatives and affricates that the acquisition of these phonemes requires an extended period of time, because speech perception and production may not be completely adult-like even at the age at which children enter school. Despite the lack of complete agreement regarding the order or the exact age of acquisition of fricatives and affricates, it is evident that the development of both the English post-alveolar fricatives and affricates extend past 4 years of age, so the acquisition of these sounds may also pose challenges to bilingual children for whom English is a target language.

1.3.2 Fricative and Affricate Development by Monolingual Spanish-Speaking Children

As it was noted previously, “standard” Spanish has three fricatives (/x s f/) and one affricate, /ʃ/ (Goldstein, 1995). Spanish also has a spirantization rule that results in voiced stops becoming fricatives except in initial and post-nasal environments and, in the case of /d/, after /l/ (Lleó & Rakow, 2004). One possible effect of a smaller Spanish fricative and affricate inventory may be that children produce these sounds earlier than their peers who are learning a more complex set of comparable sounds (such as English-speaking children). Indeed, Goldstein and Cintrón (2001) found that monolingual Spanish-speaking children appear to use fricatives more frequently at 2 years of age than their English-speaking peers, but the occurrence of a phoneme does not equal its mastery. Despite a relatively limited fricative and affricate inventory in Spanish, research on phonological development in monolingual Spanish-speaking children has also found that most fricatives and affricates are later developing sounds relative to stops and nasals (Acevedo, 1993; Eblen, 1982; Jimenez, 1987). Furthermore, mastery of some Spanish fricatives (notably, /s/) may only stabilize at around 5:7.

Linares (1981) investigated phonological acquisition in monolingual Spanish-speaking 3- to 6-year-olds in Chihuahua, Mexico and 5- to 8-year-old monolingual Spanish speaking and Spanish-dominant bilingual children in New Mexico. The author found that more than half of the consonant phonemes where acquired by 4 years of age by the children in Chihuahua and all fricatives and affricates developed between 3 and 6 years of age in monolingual Spanish-speaking children from Mexico. According to Linares (1981), /ʃ/ was acquired with 90% accuracy by 4 years of age but /s/ only by 6. Similar findings were obtained by Jimenez (1987), who conducted a study on 120 Spanish-speaking children of Mexican descent living in the United States between the ages of 3:0 and 5:7. The fricative /ʃ/ and the
affricate were produced at the very first sampling time point at 3;0 but /fl/ only stabilized with 90% accuracy at 4;3 and /fl/ reached 90% criterion level only at 4;7. The other two Spanish fricatives, /s/ and /sl/, surfaced at 3;3 and while /s/ stabilized at the 90% criterion at 4;11, /sl/ became stable only at the last sampling point at 5;7.

Eblen (1982) also found that monolingual Spanish-speaking children from Oaxaca, Mexico had the most difficulty producing /s/, considering all of the fricatives. Moreover, /s/ appeared to be more accurate in word-final than in word-initial position, as it was also found in English by Edwards (1979). The most common substitution for /s/ was the affricate (34% of the substitutions), and deletions of /s/ also occurred (28% of the time). Eblen (1982) also found that /x/ was often realized as [h] by both children and adults in Oaxaca, Mexico due to dialectal variation.

Fabiano-Smith and Goldstein (2010a) took a different approach to determining the order of acquisition of Spanish phonemes than some of the previous studies. Instead of sampling different age groups and setting an acquisition criterion, they investigated the productions of children between 3 and 4 years of age and grouped sounds as early-, middle-, and late-developing based on segmental accuracy. The authors found that /x/ was an early developing phoneme while /s/, /fl/, and /fl/ were in the middle developing group. The finding that /x/ is early acquired differs from other studies on Spanish-speaking children that generally find the voiceless velar fricative to be relatively later acquired. However, a closer investigation of Figure 2 on page 74 in Fabiano-Smith and Goldstein’s (2010a) article reveals that /x/ is the least accurate of the early-developing group; more in line with the middle-developing consonants. In fact, the accuracy of /x/ is about the same as the accuracy of /s/, a middle-developing sound. Furthermore, the Fabiano-Smith and Goldstein (2010a) article uses consonant accuracy (percentage of consonants correct) as an absolute measure of acquisition rather than a threshold as other studies did. The discrepancies between Fabiano-Smith and Goldstein’s (2010a) study and some of the previous studies may be due to methodological differences.

Despite some variation in the findings of different studies regarding the acquisition of the Spanish affricate and fricatives by monolingual children, it is clear that these phonemes are later acquired than stops or nasals, a finding that is consistent with the English results. The post-alveolar affricate seems to be used as early as 3 years of age, but it only stabilizes between 4;0 and 4;7. Moreover, the fricative /s/ only stabilizes by about 5;7. Consequently, the acquisition of fricatives and affricates by both monolingual English- and monolingual Spanish-speaking children suggests that bilingual Spanish- and English-speaking children may encounter challenges in acquiring these phonemes.

1.3.3 Fricative and Affricate Acquisition by Bilingual Children

Although similar trends in the acquisition of fricatives and affricates by monolingual Spanish- and monolingual English-speaking children have been found in that these sound classes are acquired later than nasals and stops (Anderson & Smith, 1987; Cataño, Barlow, & Moyna, 2009; Stoel-Gammon & Dunn, 1985), fricative and affricate development by monolingual children is not necessarily a predictor of how bilingual children would acquire the same phonemes. As it was noted before, the Spanish and English fricative and affricate systems differ significantly from each other (cf. Goldstein, 1995), and even those sounds that have analogs in the other language (such /fl/ or /sl/) may differ in their acoustic profile.
(e.g., Maddieson, 1980) and use (Hualde et al. 2007). Furthermore, Grosjean (1989) pointed out that a bilingual individual is not two monolinguals in one. Instead, the acquisition of multiple languages and multiple phonological systems need to be studied in their own right as well as in comparison with monolingual phonological acquisition. The current consensus on bilingual phonological acquisition seems to be that while early and proficient bilingual individuals have separate sound systems, there is also evidence for cross-language interaction (cf. Fabiano-Smith & Goldstein, 2010b; Paradis & Genesee 1996). Cross-language interaction (or interdependence) in phonology may take shape in the form of transfer, deceleration or acceleration (Fabiano-Smith & Goldstein 2010b).

Considering that both separation and interaction are attested in bilingual child phonology, one may hypothesize that learning two languages (such as Spanish and English) may result in a significant delay in phonological acquisition, because bilingual children have to acquire two different phonological systems. However, research on bilingual child phonological acquisition has, in general, not borne this out. In fact, while differences in the paths of phonological acquisition do exist between monolingual children and their bilingual peers acquiring multiple phonologies (Fabiano-Smith & Goldstein, 2010b; Goldstein & Washington, 2001), for the most part, the phonological skills of bilingual children and their monolingual peers appear to be commensurate (Gildersleeve-Neumann et al. 2008; Goldstein et al. 2005; Goldstein & Washington, 2001).

Regarding the acquisition of fricatives and affricates by bilingual Spanish- and English-speaking children, it appears that – similarly to monolingual children – bilinguals also acquire stops and nasals before affricates and fricatives (Cataño et al. 2009). Goldstein and Washington (2001) found that in the English of typically developing 4-year-old Spanish- and English-speaking bilingual children, the only two sound classes with accuracy ratings less than 90% were fricatives and affricates, and there was even more pronounced individual variation involving these sounds than normally attested in monolingual English-speaking children. However, while the order of acquisition of these sounds may be similar across bilingual children and their monolingual peers, the age at which they are mastered may differ. For example, Linares (1981) found that monolingual Spanish-speaking children from Mexico acquired /ʧ/ in a stable fashion at around 4 years of age, but their Spanish-dominant and Spanish-English bilingual peers living in New Mexico, US, had stable productions of /ʧ/ by only 5 to 8 years of age; much later than their monolingual peers from Chihuahua, Mexico. Linares (1981) attributed this difference to the children living in the US having English influences. Based on these results, it appears that bilingual Spanish- and English-speaking children display a decelerated affricate acquisition as compared to their monolingual Spanish-speaking peers.

Similar findings were obtained by Gildersleeve-Neumann et al. (2008) who compared the developing phonologies of three groups of 3-year-olds: an English only group, a primarily English group, and a bilingual English-Spanish group. The authors found no differences between the groups on vowel accuracy, but the bilingual English-Spanish group produced consonants less accurately than the monolingual English and the predominantly English groups. At initial testing, only 33% of the bilingual English-Spanish group produced affricates, as opposed to 90% of the children in the English and primarily English groups. After retesting eight months later, 100% of the primarily English group produced all fricatives and
affricates, while the bilingual English-Spanish group produced 100% of the fricatives but only 66% of the affricates. These results appear to suggest an overall deceleration in the phonology of bilingual children, but a closer examination of fricative and affricate development in bilingual children and their monolingual peers reveals a more nuanced pattern of phonological acquisition.

The accuracy with which fricatives and affricates are produced by bilingual Spanish- and English-speaking children may differ from accuracy rates attested in monolingual children, and in some cases, bilingualism may result in diminished accuracy in bilinguals, but in others, it may actually accelerate phonological acquisition. Fabiano-Smith and Goldstein (2010a; 2010b) investigated the phonological development of Spanish- and English-speaking bilingual 3- to 4-year-olds and compared them to their monolingual peers in each language. A close investigation of their data regarding the fricatives and affricates reveals that in English, bilingual children performed at about the same level as their monolingual peers on /t/ and /θ/, but the consonant accuracy of the bilinguals on /ʃ/, /ð/, and /s/ was over 10% worse, on /θ/ and /z/ over 20% worse, and on /v/ over 50% worse than the performance of their monolingual English-speaking peers (page 73, Figure 1). As for Spanish, the bilingual children were less accurate than their monolingual Spanish-speaking peers on /x/ and /s/ by about 20%, but the bilingual children outperformed their monolingual peers in Spanish on /θ/ by over 10% and on /ʃ/ by over 20% (p. 74, Figure 2). Thus, it appears that bilingual 3- to 4-year-old children match their English-speaking peers when it comes to /ʃ/ and /θ/, and they outperform their Spanish-speaking peers on the same two sounds (/θ/ and /ʃ/). These findings also support the results of Goldstein et al. (2005), who found that the English affricates were produced more accurately by primarily English-speaking children than their more balanced bilingual peers. However, fricatives were produced slightly more accurately in the Spanish of the bilinguals than in the Spanish of primarily Spanish-speaking children, albeit this varied based on the specific phoneme.

The above-reviewed literature seems to suggest that when it comes to the acquisition of fricatives and affricates, bilingual children may be generally less accurate than their monolingual peers in each language when it comes to fricative and affricate acquisition, especially in English. Nonetheless, existing research on the topic also indicates that each fricative and affricate needs to be studied in its own right, because while deceleration may characterize the acquisition of certain fricatives, others may show commensurate acquisition, and some even display acceleration. Investigating the production of /ʃ/ and /θ/ are particularly interesting, because while the affricate exists in both languages, the fricative only exists in English, and it is a later acquired phoneme, even by monolingual children. The present study addresses the production of the voiceless post-alveolar fricative in English and the voiceless post-alveolar affricate in both English and Spanish to compare the performance of bilingual children to those of their monolingual peers. Based on the existing research, three hypotheses are posited below.

**Hypothesis 1:** Bilingual Spanish- and English-speaking children will differ from their monolingual English-speaking peers in that the former group will use shorter frication durations and fricative rise times for /ʃ/ than monolingual English speakers, because Spanish does not have the stand-alone post-alveolar fricative, only the affricate that has shorter friction duration.
Hypothesis 2: Because the most likely substitute for the English /ʃ/ is the affricate, bilingual Spanish- and English-speaking children will produce more post-alveolar fricatives with a preceding burst as compared to their monolingual English-speaking peers.

Hypothesis 3: Monolingual English speakers will produce /ʧ/ with longer fricative duration than bilingual Spanish-English speakers and monolingual Spanish speakers.

2 Method
2.1 Participants
Fourteen children between the ages of 3;6 and 6;11 (years;months) participated in the present study, who were recruited from the greater Houston, Texas area (see Table 1). Based on previous studies on monolingual and bilingual children (cf. Fabiano-Smith & Goldstein, 2010a; Ingram et al. 1980; Linares, 1981), it is desirable to investigate the acquisition of post-alveolar voiceless fricatives and affricates from three to seven years of age, because these phonemes occur in three-year-olds’ productions (Jimenez, 1987) but may not be completely adult-like until much later (cf. Nittrouer, 1995). A detailed questionnaire was used to gather information on the socio-economic status and language background of each participant to determine monolingual and bilingual speech and language development history. The questionnaire also inquired about the child’s language environment, including the number of hours per day the child spoke and heard each language as well as the child’s comprehension and usage of each language. The questionnaire was largely based on Restrepo (1998) and Gutiérrez-Clellen and Kreiter (2003).

Table 1. Participant background information

<table>
<thead>
<tr>
<th>Participant Number</th>
<th>Group</th>
<th>Age</th>
<th>Gender</th>
</tr>
</thead>
<tbody>
<tr>
<td>MS001</td>
<td>Monolingual Spanish</td>
<td>3:6</td>
<td>F</td>
</tr>
<tr>
<td>MS003</td>
<td>Monolingual Spanish</td>
<td>3:6</td>
<td>M</td>
</tr>
<tr>
<td>MS005</td>
<td>Monolingual Spanish</td>
<td>4:6</td>
<td>M</td>
</tr>
<tr>
<td>MS006</td>
<td>Monolingual Spanish</td>
<td>3:8</td>
<td>M</td>
</tr>
<tr>
<td>ME010</td>
<td>Monolingual English</td>
<td>6:9</td>
<td>M</td>
</tr>
<tr>
<td>ME011</td>
<td>Monolingual English</td>
<td>3:11</td>
<td>F</td>
</tr>
<tr>
<td>ME021</td>
<td>Monolingual English</td>
<td>5:3</td>
<td>M</td>
</tr>
<tr>
<td>ME026</td>
<td>Monolingual English</td>
<td>4:10</td>
<td>M</td>
</tr>
<tr>
<td>ME027</td>
<td>Monolingual English</td>
<td>4:6</td>
<td>F</td>
</tr>
<tr>
<td>BSE001</td>
<td>Bilingual</td>
<td>6:4</td>
<td>F</td>
</tr>
<tr>
<td>BSE004</td>
<td>Bilingual</td>
<td>6:11</td>
<td>F</td>
</tr>
<tr>
<td>BSE006</td>
<td>Bilingual</td>
<td>4:9</td>
<td>M</td>
</tr>
<tr>
<td>BSE009</td>
<td>Bilingual</td>
<td>3:9</td>
<td>M</td>
</tr>
<tr>
<td>BSE010</td>
<td>Bilingual</td>
<td>5:6</td>
<td>F</td>
</tr>
</tbody>
</table>

All participants had typically developing language as determined by the parental report on the language background questionnaire and based on performing at the age appropriate level on the expressive language sub-test of the Preschool Language Scales – Fourth Edition (PLS-4) (Zimmerman, Steiner, & Pond, 2002a; 2002b). The PLS-4 is a standardized language test developed in both English and Spanish to
measure expressive and receptive language skills in children. Monolingual participants had to pass the age-appropriate section of the expressive portion of the PLS-4 in order to participate in the study, and bilingual participants needed to pass the same measure in their stronger language. None of the participants had any known physical, psychological, speech, language, or cognitive problems, as well as no significant family history of communication disorders. A pure tone hearing screening was also administered to ensure that the participants of the study had hearing within normal limits at 500, 1000, 2000, and 4000 Hz at 25 dB hearing level. In addition, Spanish-speaking children were selected for participation only if they spoke a dialect of Spanish other than from Chihuahua, Mexico due to its allophonic usage of [ʃ] for /ʧ/ (Hualde et al. 2007).

Participants belonged to one of three groups: monolingual English (ME), monolingual Spanish (MS), and bilingual English-Spanish (BES) as determined by the language background questionnaire. The ME group and the BES group each had 5 participants and the MS group had 4 participants (see Table 1). A child was considered functionally monolingual if s/he used one language (Spanish or English) over 80% of the time, and a child was considered bilingual if both languages were used at least 20% of the time. A minimum of 20% use criterion in each language was chosen based on work done by Pearson, Fernandez, Lewedeg, and Oller (1997). An additional criterion specified that in order to be considered bilingual, a child also needed to be able to produce full sentences in both languages spontaneously, which served as an objective measure in addition to parental language use reports to ensure that the bilingual participants were competent in both of their languages (see also Bunta & Ingram, 2007).

2.2 Materials

A picture naming task was used to elicit single words with the target phonemes in initial position as well as foils in both English and Spanish. Black and white line drawings appropriate for young children that did not demonstrate a preference to one particular culture were presented via a laptop computer to the children, one image at a time. There were 60 items for Spanish and 90 for English, each representing words commonly occurring in the vocabularies of young children. The age-appropriateness of the target items was ensured by choosing words that have been demonstrated to be amenable to elicitation from preschool-age children in previous studies (e.g., Fabiano-Smith & Goldstein, 2010a; Hase, Ingram, & Bunta, 2010). Due to ensuring the age-appropriateness of the items, we did not encounter significant problems in eliciting the words from the children. The phonemes that were targeted in the activity (the post-alveolar voiceless fricative and affricate) occurred in word-initial singleton position (e.g. “chair”) in a syllable that had primary stress. Each child was asked to independently name each picture by answering the question, “What is this?”

If the child could not label a picture, the investigator used delayed imitation to elicit the target word by stating, “This is a _____. What is this?” All of the children were able to produce the targeted items with the above-described prompting procedure. All samples were audio recorded digitally onto a laptop computer with an Echo Indigo IO external sound card at 44 kHz and 16 bits. A Sennheiser EW300 G2 wireless lavalier microphone system was used to capture the sound files. The microphone was positioned about 8 inches from the participants’ mouths to ensure consistency across the samples.
2.3 Procedure

The data were collected in a separate room of the preschool center the child attended, in the child’s home or in the Bilingual and Cross-Linguistic Language Laboratory in the Department of Communication Sciences and Disorders at the University of Houston. In order to limit code-mixing and code-switching, the bilingual participants were sampled on two separate days with each language sampled on a separate day by different experimenters. Each experimenter had experience working with children and was aware of culturally appropriate ways of interacting with bilingual preschool-age children. In addition to receiving parental consent prior to beginning the language sample, all children provided verbal assent to participate in the study. Upon granting written parental consent and child verbal assent, each child participated in a hearing screening, followed by a language screening via the PLS-4. Afterwards, the picture naming task was administered, followed by the story tell task (the latter was not analyzed for the present study). The entire procedure took approximately 45 minutes for monolingual speakers and two 45-minute sessions for bilingual speakers (one in Spanish and one in English).

All of the single-word samples were transcribed phonetically by the second author, and about 30% of the transcriptions were verified by another transcriber. All transcribers were trained in using the International Phonetic Alphabet (IPA) and had prior experience transcribing children in the language in which they transcribed. The existence of a burst or the lack thereof was determined auditorily and confirmed acoustically based on the time waveform and the spectrogram.

Acoustic analyses were conducted to measure the relevant properties that differentiate the targeted sound segments, /ʧ/ and /ʃ/. As it was reviewed in the introduction, post-alveolar fricatives and affricates are differentiated based on the presence or absence of a silent period (affricates have a gap and fricatives do not), the existence or lack of a burst (affricates have a burst), and the entire duration and the duration of the rise time of the fricative portion (fricatives having much longer overall durations and longer rise times than affricates). In the present study, we investigated the last three cues (the burst, the duration of the frication, and the rise time of the frication) but not the silent period, because the gap cannot be reliably measured for voiceless fricatives versus affricates in word-initial position.

Wavesurfer (Sjölander & Beskow, 2010) was used to conduct the acoustic measurements using a time waveform and a wide-band spectrogram. The settings for the spectrographic display were the following: 1024 window length, 350 – 400 Hz bandwidth (depending on the child’s fundamental frequency), 0.8 pre-emphasis, and a display with a range of 10000 Hz. First, as it was previously noted, the time waveform and the spectrogram were used to determine whether or not there was a burst. The second analysis involved measuring the duration of the entire fricative and the fricative portion of the affricate. The third aspect measured was rise time of the fricative (in the fricative itself and the fricative portion of the affricate). Regarding the duration measurements, a 10-millisecond window of error was accepted as the same duration measurement as it is commonly done in research involving segmental duration measurements (e.g., Peterson & Lehiste, 1960). Over 700 acoustic measurements were conducted and verified by two experimenters. Figure 1 illustrates the measurement of the fricative portion of an affricate. The area highlighted in yellow demonstrates the duration measurement.
2.4 Reliability calculations

As noted above, all participant samples were phonetically transcribed by the primary investigator and 30% of the data were transcribed by a research assistant trained in phonetic transcription to ensure inter-judge reliability. Inter-rater agreement on phonetic transcriptions exceeded 90%. Regarding the burst and the duration measurements, 100% of the sample was verified by two individuals trained in acoustic analysis. Inter-rater agreement exceeded 95% on all duration measurements, allowing for a 10-millisecond error window, as described above.

3 Results

According to Hypothesis 1, bilingual Spanish- and English-speaking children were expected to have shorter fricative durations and rise times than their monolingual English-speaking peers. The results only partially supported this hypothesis. In comparing the bilingual participants’ friction duration and rise time of /f/ to those of their monolingual English-speaking peers, no significant differences were found on average fricative durations. Even though both the duration and the rise time averages of /f/ were longer for the monolingual English-speaking participants as compared to those of their bilingual peers speaking English, these differences were not statistically significant. For fricative durations, the bilingual children speaking English had a mean length of 164 msec (SD = 0.025) while their monolingual English-speaking children’s fricatives were 195 msec (SD = 0.062), a difference that was not statistically significant (t (8) = -1.04, NS, Cohen’s d = -0.73). The same pattern was observed for fricative rise times: the average for the bilingual children was 89 msec (SD = 0.023) and the monolingual English-speaking children’s average was 115 msec (SD = 0.035), a difference that was also not statistically significant (t (8) = -1.381, NS, d = -0.98). It is worth noting, however that the effect sizes for both of these contrasts were sizeable, so further investigation revealed that on the word “shovel”, both the duration of the fricative and the rise time of the fricative were different (t (8) = -2.320, p = 0.049, d = -1.64 for the duration of the initial fricative in “shovel”, and t (8) = -2.926, p =0.019, d = -2.07 for the rise time of the fricative). The lack of statistical precision may have been due to the small sample size, but the large effect sizes are promising, especially because statistical
significance alone may not be sufficient to capturing existing differences (see Ziliak & McCloskey, 2008 for a discussion on effect size versus precision). Figures 2 and 3 display the average durations and rise times of the voiceless post-alveolar fricative produced by bilingual and monolingual children speaking English along with the related standard errors for the means.

**Figure 2.** Durations of the English /ʃ/ with standard error of the mean

**Figure 3.** Rise times of the English /ʃ/ with standard error of the mean

Our second hypothesis predicted that bilingual Spanish- and English-speaking children will produce more post-alveolar fricatives with a preceding burst than their monolingual English-speaking peers. Qualitative analyses did support this hypothesis. All of the bilingual children had at least one of their word-initial English /ʃ/ phonemes produced with a preceding burst while only one of the monolingual English-speaking children produced a burst for two of her target words. In all,
bilingual children produced a burst before 15 of their fricatives, whereas only one monolingual English-speaking child produced 2 fricatives with a preceding burst. It also appears that some words prompted more tokens of affrication than others: “sheep” and “shovel” had 4 tokens of affrication each, “shark” had 3, “shot” had 2, and “shovel”, “shirt”, “shadow”, and “shell” had 1 each. The only word that did not prompt affrication in the bilingual children’s productions was “shoe”, possibly because it has a meaningful minimal pair in “chew”. In the discussion, we return to the issue of variation based on lexical items. Figures 4 and 5 provide examples of the word “shirt” produced by two different bilingual children. Figure 4 displays a typical, non-affricated production of /ʃ/, and Figure 5 illustrates an affricated fricative produced by one of our bilingual participants.

![Figure 4](image1.png)

*Figure 4. Non-affricated example of /ʃ/ by a bilingual child in the word “shirt”*

![Figure 5](image2.png)

*Figure 5. [tʃ] for /ʃ/ substitution by a bilingual child in the word “shirt”*

The third hypothesis predicted that the duration of the fricative in affricates will differ in the productions of monolingual English-speaking children, their bilingual peers, and their monolingual Spanish-speaking peers. Our findings did not support the third hypothesis. There were no statistically significant differences between the
bilingual English speakers and their monolingual English-speaking peers on the duration of the fricative portion of the affricate (t (8) = -0.81, NS, d = -0.57) or the rise time of the fricative (t (8) = -0.461, NS, d = -0.33), and the effect sizes were in the small to medium range. The comparison of the fricative duration of the affricate between monolingual Spanish and monolingual English speakers did not reveal any statistically significant differences either, but there was a large effect size (t (7) = -1.78, NS, Cohen’s d = -1.35). Thus, the first hypothesis was rejected, but the effect sizes are worth noting (see also Ziliak & McCloskey, 2008). Figure 6 illustrates the fricative duration of the voiceless post-alveolar affricate for the participant groups. These, and other differences involving the voiceless post-alveolar affricate will be discussed in the next section.

![Fricative Duration of Affricate](image)

**Figure 6.** Fricative duration of the voiceless post-alveolar affricate

### 4 Discussion

The main purpose of the present study was to investigate the production of the voiceless post-alveolar affricate in English and Spanish and the voiceless post-alveolar fricative in English by bilingual children and their monolingual peers. Our results indicate that specific cross-language effects were evident in the bilingual children’s productions. Most notably, our results indicate that affrication of the voiceless post-alveolar fricative is more frequent (albeit not predominant) in the speech of bilingual Spanish- and English-speaking children as compared to the productions of monolingual English-speaking children. Furthermore, while there was not an overall statistically significant difference between the duration of /ʃ/ in the productions of bilingual versus monolingual children speaking English, effect sizes appear to indicate that, in general, monolingual English-speaking children tend to have longer /ʃ/ fricatives than their bilingual peers. Figure 7 illustrates the differences regarding the duration of word-initial /ʃ/ produced by bilingual and monolingual children speaking English in order of the largest effect (leftmost pair of bars) to the smallest one (rightmost pair). The largest effect size was observed for the word “shovel” (d = -1.64) and the smallest one for “shower” (d = 0.004). It is also interesting to note that all of the large and even the medium effect sizes were
supporting the direction of monolingual English speakers having longer fricatives than their bilingual peers. The only exception to this was the word “shoe” that displayed a relatively large-medium effect \((d = 0.77)\) with the bilinguals having longer fricative durations than their monolingual peers. Nevertheless, the general trend was towards monolingual English-speaking children having consistently longer voiceless post-alveolar fricatives (see Figure 7).

It is also important to note the results varied considerably depending on the target word; a finding that is consistent with Ingram et al.’s (1980) results, who found that the target word had a significant impact on how monolingual English-speaking children produced word-initial fricatives and affricates. Apparently, this word-based variability applies to both monolingual English-speaking children and their bilingual English- and Spanish-speaking peers when it comes to fricative production.

Comparing the production of the affricate across the groups did not yield any statistically significant differences. However, the effect size of the comparison between monolingual English and monolingual Spanish speakers was very large, suggesting that monolingual English speakers may have longer fricative portions in their affricates than their monolingual Spanish-speaking peers. This observation is consistent with the results of previous research that found differences between Spanish and English fricative durations in affricates produced by monolingual adult speakers (Maddieson, 1980). In addition, in comparing the fricative portion of the affricate produced by monolingual English speakers and their bilingual peers speaking English, we found that the effect size was medium. Figure 6 illustrates the trend by which monolingual English speakers tend to have the longest fricative portions in the affricate, followed by bilinguals, and then by monolingual Spanish speakers. When we contrasted the English and Spanish fricative portions of the affricate, it seems that bilingual children use the same values for both languages; thus, they may not differentiate their Spanish from their English /ʃ/ as indicated by a negligible effect size \((t(4) = 0.387, \text{NS}, d = 0.10)\). The lack of differentiation of /ʃ/
in the English and Spanish of the bilingual participants is an example of bilingual phonological interaction, as it was found by Fabiano-Smith and Goldstein (2010b).

Previous studies also investigated the relevance of both fricative durations and rise times in distinguishing post-alveolar affricates from fricatives. While some studies found that rise times were relevant to differentiating between fricatives and affricates (e.g., Gerstman, 1957) others have found duration to be more reliable (Kluender & Walsh, 1992; Maddeison, 1980). Non-parametric Spearman's $\rho$ correlation coefficients revealed that fricative duration and rise time are highly correlated and may be redundant with each other. The rise time and duration of /ʃ/ was correlated at $\rho = 0.83$ at $p = 0.003$. For /tʃ/, the correlation between the duration of the fricative portion and its rise time was $\rho = 0.86$ at $p < 0.001$. The duration of the initial voiceless post-alveolar affricate including the burst was also highly correlated with the fricative portion of the affricate, for obvious reasons, since the burst only added a few milliseconds to the duration of the affricate ($\rho = 1.00$ at $p < 0.001$). Thus, based on our results, rise times alone do not seem to add information above and beyond what fricative duration can provide.

The present study is not without its limitations. Considering that many of the contrasts in our study had large and medium effect sizes while few of those comparisons were statistically significant, it appears that the number of participants in the current study was insufficient. Having more participants would increase statistical power and may lead to uncovering more effects that could be found with a limited sample. Furthermore, the age range and of the participants varied within and across the groups, and it should be more controlled more carefully in future studies. Building on this pilot work, future studies should control the participant characteristics as much as possible.

There was also considerable variability in individual scores and based on different words. However, variability is very common in bilingual and monolingual child language studies, and rather than trying to limit variability in child phonology studies, variation deserves further investigation. In fact, future studies should include a larger range of lexical items precisely so that fricative and affricate development can be studies in a more comprehensive fashion.

Future studies on bilingual phonological development should also include both phonetic transcriptions and other (e.g., acoustic) analyses to gain a more complete and accurate picture of children’s phonological skills. As Edwards and Munson (in press) note, phonetic transcriptions by native speakers include listener bias and tolerate various levels of distortion, so that even some child non-adult-like productions may be transcribed as correct because they often fall within the adult perceptual space. Thus, while phonetic transcriptions are still the mainstay of phonological analyses, tracking subtle changes and progress over time with transcription-based analyses alone may be insufficient. Acoustic analyses may accompany and complement phonetic transcriptions, because a combination of the two can provide a fuller picture of phonological development and enables researchers to capture and track phonological development more accurately than either one alone, especially when it comes to the phonological skills of bilingual children.
5 Conclusion
The present study contributes valuable information to our understanding of the acquisition of voiceless post-alveolar fricatives and affricates by bilingual Spanish- and English-speaking children and their monolingual peers in particular, but the findings also have significant ramifications for bilingual phonological development, in general. Overall, bilingual Spanish- and English-speaking children produced more voiceless post-alveolar fricatives with an initial burst and with shorter durations than their monolingual English-speaking peers. A general trend regarding fricative durations in /tʃ/ appeared to indicate that monolingual English speakers had the longest fricatives in the affricate, followed by bilinguals, and monolingual Spanish speakers had the shortest ones. Furthermore, the fricative durations in /tʃ/ in the Spanish and English of the bilinguals were not differentiated, suggesting cross-language interaction. There was also considerable individual variability as well as variability based on the target words. Finally, post-hoc analyses indicated that the fricative rise time and duration may be redundant features in the productions of bilingual and monolingual children. This pilot study obtained promising results that future research should continue to expand on and explore. These findings should not only be verified by future research, but this line of research will lead to future discoveries and has important ramifications for bilingual phonological development and phonological acquisition theory.

6 Acknowledgements
We would like to thank Sarah Foreman, Michelle Norton, Leah Sheaffer, Elcia Olivio, and Cynthia Peña for their assistance with data collection and analysis. We also express our gratitude to Rachel Aghara and Elizabeth Goodin-Mayeda for their feedback on an earlier draft of this paper. The faculty and staff at Los Niños Community Family Center, Precious Jewels Alpha Learning Center, and M.D. Anderson YMCA Preschool were also instrumental in making this project a reality by assisting us with access to their students and facilities. Finally, we thank all of the children and the parents who chose to participate in the present study. Thank you.

References


Shriberg, L. D. 1993. Four new speech and prosody-voice measures for genetics research and other studies in developmental phonological disorders. Journal of Speech and Hearing Research 36, 105-140.


A CASE STUDY ON THE EFFECT OF DISCOURSE TYPE ON FLUENCY LEVEL IN STUTTERING

Mária Gósy¹,² and Judit Bóna²
¹Research Institute for Linguistics, Hungarian Academy of Sciences, Budapest, Hungary
²Department of Phonetics, Eötvös Loránd University, Budapest, Hungary

E-mail: gosy.maria@nytud.mta.hu, bona.judit@btk.elte.hu

Abstract
The present study analyzed the effects of five different discourse types (narrative, conversation, recall, reading, sentence repetition) on the fluency of a 56-year-old Hungarian-speaking, right-handed female who stutters. The occurrences (in percentage) and durations (in ms) of filled pauses, word and part-word repetitions, and prolongations as well as the cutoff-to-repair intervals were analyzed by means of Praat software. The results confirmed that the analyzed disfluencies showed discourse type dependency where reading, sentence repetition and narrative formed one category, while recall and conversation formed another, although the connection between them is not as close as that between the discourse types in the former category. Recall seemed to be an independent category demonstrating that almost all analyzed disfluencies were either the most frequent ones or showed the longest durations in this discourse type. The interrelations of various speech planning levels were discussed.

1. Introduction
Stuttering is a multidimensional/multifactorial speech disorder that – as it is widely accepted – affects the fluency of speech (Burger and Wijnen, 1999; Curlee, 2004; Pochman and Attanasio, 2007; Craig, Hancock and Craig 2009, etc.). Although on the surface it is the involuntary fluency problem that is its most conspicuous feature, it is exactly this same feature that seems to be the most controversial (e.g., Perkins, 1995; Howell, Sackin, Glenn and Au-Yeung, 1997; Bortfeld, Leon, Bloom, Schober and Brennan, 2001; Ginsberg, 2000; Yairi, 2007). There are a variety of other factors that may influence stuttering. These are the communication situation, familiarity with the speech partner, content of the message, the linguistic complexity of the texts, emotional condition, etc. (e.g., Wall, Starkweather and Cairns et al., 1981; Bosshardt, 1997; Vasic and Wijnen, 2005; Howell, Au-Yeung, Yaruss and Eldridge, 2006). These factors have different effects on different speakers. There can be interactions among the emotional, cognitive, and linguistic factors, which in turn interact with the speech motor system (e.g., Maner, Smith and Grayson, 2000; Irwin, 2006; Blomgren and Goberman 2008; Smith, Sadagopan, Walsh and Weber-Fox, 2010).

Persons who stutter (PWS) can be recognized because they have relatively frequent disfluent episodes and specific blocks in their speech flow. Stuttering includes repetitions of sounds and/or syllables, part- or whole words, prolongations of speech sounds, avoidance of words, substitutions, or blocking of sounds
(Bloodstein, 1993; Craig, Hancock, Chang Chang, McCready, Shepley and McCaul, 1996; Craig, Hancock and Craig, 1996). The ICD (2007) defines stuttering as speech that is characterized by frequent repetition or prolongation of sounds or syllables or words, or by frequent filled pauses or pauses that disrupt the rhythmic flow of speech. It should be classified as a disorder only if its severity is such as to markedly disturb the fluency of speech.

Many disfluencies pass unnoticed or are accepted as speaker-specific characteristics in persons with normally fluent speech (NFS), although blocks and frequent part-word repetitions are conspicuous and less acceptable to listeners (Roberts, Meltzer and Wilding, 2009). Some researchers distinguish PWS from NFS on the basis of the frequency of certain types of disfluencies (Postma and Kolk, 1993). However, others claim that it is unlikely that PWS can be distinguished from NFS based solely on the number of disfluencies (Wingate, 1988; Bloodstein, 1993; Parry, 2009). As has long been known, there are many NFS people who are disfluent from time to time, but that does not make them stutters, since blocks are extremely rare with them (Roberts et al., 2009). PWSs’ speech contains a considerable amount of disfluency that also varies substantially and includes long fluent phases as well.

It has been proposed that stuttering is the result of covert detection and correction of errors in the articulatory plan (Postma and Kolk, 1993; Brocklehurst, 2008). Some researchers, while sharing this basic assumption, regard stuttering as a monitoring deficit (Lieshout, Peters, Starkweather and Hulstijn, 1993; Kolk and Postma, 1997; Vasic and Wijnen, 2005). PWS seem to detect and repair their covert errors with oversensitivity. This means that they are ready to correct possible errors before they are actually manifested. In other words, PWSs are generally able to realize that a particular word is going to cause trouble (and try to avoid pronouncing it) and so they are ready to correct even non-existing errors. Self-monitoring on the surface is signaled by two facts: halting of speech and the cutoff-to-repair interval (Levelt, 1989; Hartsuiker and Kolk, 2001; Brocklehurst, 2008). The cutoff-to-repair interval is manifested through filled or unfilled pauses providing an opportunity to plan repair. The cutoff-to-repair interval is optional, but its duration provides information about the time needed for controlling and correction of errors.

There are several models suggesting that stuttering can be understood from a multidimensional perspective, like the dynamic multifactorial models (Smith, 1997; Smith, 1999; Smith and Kelly, 1997), The Demands and Capacities Model (DCM: Starkweather et al., 1990; Kelly, 2000) or the Integrated Multidimensional Model of Stuttering (CALMS: Healey et al., 2004). Since a PWS may react to different communication situations in different ways, the processes taking place in the speech motor system may be influenced by the interacting factors, depending on the discourse types occurring in a given situation. In order to obtain objective data on the effects of particular discourse types, a PWS’s, speech must be studied under natural circumstances (Tetnowski and Damico, 2001).

Spontaneous and non-spontaneous speaking tasks require different contributions from the different levels of speech planning. If an utterance that is going to be pronounced does not have to be formulated either semantically or grammatically, as it is the case in reading (Max and Baldwin, 2010), then only the lower levels are activated. Phonetic planning and execution in reading require the decoding of the written letter strings. The PWS’s main problem here is that some of the speech sounds s/he has to pronounce (e. g., stops or nasals) and/or some (long) words might
give her trouble. In addition, she cannot avoid certain phonologically complex words (Smith et al., 2010). Therefore, even though articulation planning may be perfect, execution may fail in some cases because of the speaker’s excessive monitoring and fear of forthcoming trouble.

The speech planning processes are similar in sentence repetition; however, this task requires attention to remembering the heard sentences (Bajaj, 2007b). In this task the PWS encounters not only speech sounds and words that might be difficult for her to articulate, but s/he also has to remember the phrases heard until the articulation gestures are completed. Monitoring has to control the phonetic plan and its execution on the one hand, and controlling the exact repetition of the sentences heard according to their morphological and syntactical structures, on the other hand.

In a narrative, all of the processes of the speech planning mechanism must be activated. The speaker has to cope with his/her own thoughts and has to select those that s/he intends to share with the interlocutor. The selection of thoughts, their formulation into grammatical forms, phonological and phonetic planning, as well as pronunciation are in progress almost at the same time (e.g., Levelt 1989; Bajaj, 2007a). Speech production planning is supposed to focus, in this case, on the higher level processes rather than on those at the lower levels. The operations at higher levels of encoding are thought to be responsible for some difficulties in the PWS’s speech (e.g., Ratner, 2005).

Spontaneous speaking during a recall task (e.g., Scott et al., 1995; Polyna et al., 2009) seems to require all those processes that contribute in the other four discourse types. The speaker has to interpret the story s/he has heard. Memory traces have to be recalled successfully in order to retell the story and to use appropriate grammar and vocabulary (as in the other two spontaneous discourse types). Self-monitoring also requires permanent comparison of what the speaker has recalled with the facts of the original story.

Conversation, as a discourse type, is similar to narrative: all of the processes of the speech planning mechanism are active (e.g., Logan and Conture, 1995). However, the speaker – as one of the participants in the speech situation – has to co-operate with the other speakers. Although a speaker in a conversation is not expected to speak continuously for a long time, as happens in a narrative, s/he has to integrate his/her own thoughts with the reactions of other speakers. The speaker has to observe the flow of conversation and has to make plans when and how to join in or interrupt ongoing speech. Self-monitoring is responsible for all these processes, including subsequent corrections relating to content, grammar or pronunciation.

In sum, the speech planning characteristics of different discourse types suggest that they might have an effect on the fluency level of a PWS’s speech. We assume that this effect can be identified by analyzing various types of disfluencies in speaking. Spontaneous speech is full of disfluencies, such as filled pauses, repetitions, false starts, prolongations, etc., signaling the speaker’s difficulty during speech planning (e.g., Levelt, 1989; Fox Tree, 1995; Shriberg, 2001; Bortfeld et al. 2001; Gósy 2003; Watanabe et al., 2008).

Our understanding of stuttering implies that the speech task required in different discourse types has a significant impact on the PWS’s speech, according to the multifactorial models (mentioned above). We think that this impact is shown on the surface by the diverse numbers of occurrences and diverse durations of disfluencies, which are dependent on discourse types. A limited number of studies suggest
discourse type dependent characteristics of the disfluencies mainly in children who stutter (e.g., Trautman et al., 1999; Logan et al., 2011). Results of a recent study with PWS from neurogenic origin supported the effects of various speaking tasks on their speech (Tani and Sakai, 2011).

The purpose of this study was to investigate whether or not the characteristics of four types of disfluencies in a Hungarian-speaking PWS were dependent on discourse type. Filled pauses, prolongations, part-word repetitions and word repetitions were analyzed in a PWS’s speech – disfluencies that also occur in NFS’s speech. Our intention was to learn whether the various discourse types (narrative, recall, conversation, reading and sentence repetition) affect the occurrence and temporal properties of the four analyzed types of disfluencies. In addition, we wanted to obtain information on possible differences in the control processes for part-word repetitions and word repetitions by analyzing the durations of their cutoff-to-repair intervals in various discourse types.

We hypothesize that the occurrence and temporal patterns of the disfluencies in the PWS’s speech will be affected by the different speaking tasks presented by the analyzed discourse types. We assume that the duration needed for control in part-word repetitions and word repetitions would also be discourse type dependent. The importance of such a study lies in its ability to provide measured data on the effect of discourse types on a PWS’s speech.

2 Subject, method, material

2.1 Subject

In this paper we report on a 56-year-old, Hungarian-speaking, monolingual, right-handed female PWS with a history of clinically diagnosed developmental stuttering. She started stuttering at the age of 3. Up to that age her first language acquisition was typically developing. She was a severe stutterer according both to the Hungarian Classification Scale of Stuttering and the Stuttering Severity Instrument SSI-3, and she regards herself as such (Lajos, 2003; Riley, 1994, respectively). She had no hearing, neurological, mental, speech or language deficits (other than stuttering). She had no reading problems.

2.2 Speech material and measurements

A total speech sample of 1.2 hours was audio recorded with the subject in a sound proof chamber at the Phonetics Laboratory of the Research Institute of the Hungarian Academy of Sciences. For recordings, a unidirectional high-quality microphone and Goldwave software connected to a computer were used. The speech samples were recorded following the Spontaneous Speech Corpus of Hungarian (BEA) protocol. All the discourse types involved presented the speaker with a different task. From each spontaneous discourse type, we selected a sample that was approximately 10 minutes long and began 8 minutes after the start of the recording. The speech sample was representative of the patient’s usual speech pattern (see Van Borsel and Taillieu, 2001). The measured speech samples contained: 1961 syllables (553 words) in narrative discourse, 2164 syllables (571 words) in conversation, 2186 syllables (597 words) in recall, 1287 syllables (243 words) in reading and 1181 syllables (213 words) in sentence repetition. Hungarian has many long words in spontaneous speech as a result of its agglutinative character (frequently up to 7 or more syllables with a mean of about 3.5 syllables).
The discourse types were narrative (the participant was asked to speak about her life, family, work and hobbies); recall of an orally presented story (the participant was told a story and was asked to summarize what she had heard); conversation (only the participant’s speaking turns were considered; the topic was job openings for young graduates); sentence repetition (22 well-formed sentences of various lengths were compiled for repetition, containing 9 to 15 words / 19 to 36 syllables); and reading (the participant was asked to read aloud a popular scientific text without rehearsal).

The durations of 913 filled pauses (in other words hesitations), 287 prolongations, 215 word repetitions and 335 part-word repetitions were found in the speech sample and analyzed. The two authors encoded and measured the disfluencies separately and, in cases of rare disagreement, two other phoneticians were consulted. Filled pauses were defined as disfluencies occurring between two different words with or without a preceding and/or following unfilled pause. Prolongation arose with vowels and also with consonants, particularly if they were continuants. The occurrences of all these disfluencies and the durations of filled pauses and prolongations were analyzed. In the case of repetitions, the durations of words in both the first and the second production were measured. The duration of cutoff-to-repair intervals in word repetitions and part-word repetitions were measured. Vowel duration was measured between the first and last glottal pulses of the vowels, while that of the consonants was measured according to their acoustic structure. The duration of the cutoff-to-repair intervals was measured from the interruption point to the onset of repair. All the measures were conducted across both the fluent and the stuttered periods of the speech samples (without any selection).

The digital recordings were submitted to acoustic-phonetic analysis (Praat software: Boersma and Weenink, 2004) using a 44.1 kHZ sampling rate with a 16-bit resolution. To test statistical significance, analysis of variance (one-way ANOVA), Tukey’s post-hoc tests, the Kruskal-Wallis test, and hierarchical cluster analysis were used, as appropriate (SPSS 14.0). The confidence level was set at the conventional 95%. Hierarchical clustering is a type of cluster analysis based on the assignment of a set of observations to subsets called clusters.

3. Results

The five different discourse types used in the research involved different communication tasks, and consequently different speech planning processes and probably different types of self-monitoring. In sentence repetition and reading the speaker was presented with semantically and syntactically ready-made sentences, so it was only articulatory planning and execution that required attention. This means that her task was simplified compared to spontaneous talking. The tasks of telling a narrative, taking part in a conversation and recalling a story required the use of higher speech planning levels. These are more complex tasks than reading and repetition. The two examples below have been selected from the narrative task: the first one (1) exemplifies a relatively disfluent production, while the second one (2), a relatively fluent episode within the speech sample. The various types and forms of disfluencies have been marked in bold. Prolongations are indicated by double letters, and the letters ö, öm, ööm, öhm demonstrate filled pauses.
The more disfluent episode contains filled pauses of various forms, word repetition (hogy hogy ‘that that’), part-word repetitions (s s sport) and prolongations (bizzonyos).

This other episode is similar to a NFS’s speech sample. It contains an ö type filled pause and a word repetition (akkor akkor ‘that time that time’).

3.1 Filled pauses

Filled pause is one of the most frequent disfluencies in spontaneous speech: its primary function is to provide time to surmount difficulties in speech planning (e.g., Shriberg, 2001; Gósy, 2003; Watanabe et al., 2008). In Hungarian, filled pauses in most cases involve the use of a vowel-like sound of varying length close to [ø] or the neutral vowel [ə] (see Horváth 2010). Most of the 913 filled pauses were articulated with this neutral vowel. The frequency of filled pauses showed remarkable differences across discourse types. They occurred rarely in reading (5.9/minute) and in sentence repetition (10.5/minute), while considerably more frequently in conversation (20.1/minute), in narrative (26.7/minute) and in recall (27.4/minute).

By way of comparison, an average of 3.82 filled pauses per minute was found in Hungarian-speaking NFSs, with wide individual differences ranging from 0.8 to 9.5 per minute.

The mean duration of all filled pauses was 309 ms which is almost the same as that found with NFSs (307 ms on average, see Horváth 2010). The longest filled pauses were found in recall (mean: 349 ms, SD = 255.95), while shorter ones occurred in sentence repetition (mean: 312 ms, SD = 180.89), in narrative (mean: 294 ms, SD = 205.42), in conversation (mean: 284 ms, SD = 210.82) and in reading (mean: 221 ms, SD = 148.12). Since the data were not normally distributed, they were transformed into a logarithmic scale. One-way ANOVA was used on these data, and statistical analysis revealed significant differences depending on discourse type ($F(4,912) = 3.436$, $p = .008$), see Fig. 1. The post-hoc Tukey test showed significant differences between recall and the three other discourse types, reading ($p = .019$), narrative ($p = .043$) and conversation ($p = .021$).
Figure 3.1. Durations of filled pauses across discourse types (median and range). (R = reading, SP = sentence repetition, RC = recall, N = narrative, C = conversation.)

The occurrences and the durations of filled pauses showed a close relationship in our PWS’s spontaneous speech samples. The more frequent the filled pause, the longer its duration (e.g., in recall) and vice versa; the less frequent its occurrence, the shorter its duration (e.g., in conversation).

3.2 Prolongations

Although this type of disfluency is regarded as one of the primary symptoms of stuttering, it occurs in NFSs’ speech as well, and the occurrence of various prolongations here is about 1.5 incidents per minute in Hungarian speech (Bóna and Imre, 2010). The occurrence of prolongations in our PWS’s speech samples was most frequent in recall (9.22 per minute) and in sentence repetition (8.8 items per minute), while it occurred less frequently in reading (6.37 per minute) and in narrative (5.29 per minute). The least frequent occurrence was found in conversation (3.1 per minute).

Although several consonants were involved in the prolongations, most prolonged speech sounds involved vowels, particularly those in the definite article (a/az). Where prolongation occurred, it affected the initial speech sounds of words: in 95.5% of all cases in recall, and in 93.4% of all cases in conversation. Prolongation occurred on the first speech sound in 84.2% of all cases in sentence repetition, in 75% of all cases in narrative and in 67.7% of all cases in reading. The next most frequently prolonged speech sounds turned out to be word-final sounds (11.18% of the total). The most frequently prolonged last sounds were found in reading (22.6%), then in narrative (13.5%) and somewhat more rarely in sentence repetition (10.5%). The occurrence of last sound prolongation decreased significantly in conversation (6.6%) and in recall (2.7%). Prolongations rarely appeared on the second or third sounds of words (about 11% of all incidents in all discourse types) while somewhat more frequently on sounds close to the end of words (about 17%).

Prolongations appeared in content words in 41.0% of all cases in narrative, in 16% in recall, in 38.7% in reading, in 33.9% in sentence repetition, and only 3.3% in conversation. The proportions of prolongations occurring on content words in spontaneous speech samples were lower in our PWS (21% on average) than those
obtained with NFSs, where prolongations in content words averaged 51% (Bóna, 2008). Prolonged speech sounds seem to be more characteristic of narrative, reading and sentence repetition compared to recall and conversation in our PWS’s speech.

The durations of prolongations seem again to depend on discourse type (in reading: 355 ms, SD = 134.83, in sentence repetition: 402 ms, SD = 203.40, in recall: 402 ms, SD = 203.40, narrative: 314 ms, SD = 111.37 and in conversation: 467 ms, SD = 209.91). The longest prolonged speech sounds occurred in conversation, while the shortest ones in the narrative. The durations of prolongations exceeded 1000 ms in recall, in conversation and in sentence repetition (the longest prolongation was 1210 ms in conversation, 1346 ms in sentence repetition and 1456 ms in recall). There were no significant differences in the prolongation durations across discourse type. The duration of the prolonged sounds in NFSs’ spontaneous speech ranged between 150 and 600 ms, with a mean value of 320 ms (Bóna, 2008). These data correspond to our PWS’s values in narrative (mean: 314 ms, SD = 111.37).

The interrelations between occurrence and duration showed that prolongations were least frequent in conversation, but their durations were the longest here. They were both relatively long and frequent in narrative, while relatively less frequent and shorter in recall. Sentence repetition and reading occupy an intermediate position concerning the interrelations of occurrence and duration (Fig. 2).

![Figure 3.2. Durations of prolongations across discourse types (median and range).](image)

(R = reading, SP = sentence repetition, RC = recall, N = narrative, C = conversation.)

### 3.3 Word repetitions

Howell (2007) divides fluency failures into two classes (called ‘stalling’ and ‘advancing’) where stallings are characterized by repetition of one or more words. He claimed that repetitions were usually function words and that repeated simple words often preceded more complex words (Howell, Au-Yeung and Sackin, 1999; Howell and Dworzynski, 2005). Our data fully supported this claim: there was only one content word repeated – the word legfontosabb ‘most important’ – in the whole of our speech material. The proportion of function words in repetitions was 99.75%. The dominance of function words might also be attributed to the (already mentioned) fact that content words are relatively long in Hungarian and the speaker does not
want to spend extra time repeating long words (function words contain generally one or two syllables). Besides, in non-stuttered speech, the proportion of function words in repetitions is also high, 92.9% (Gyarmathy, 2009).

The occurrence of repetition is heavily dependent on discourse type. There were almost no repetitions in reading (1.08 per minute) and in sentence repetition (0.74 per minute). The increase in the number of repetitions in the spontaneous speech samples was evident: it was 5.07 per minute in recall and 5.87 per minute in narrative. The highest frequency of repetition was found in conversation (7.83 per minute). For comparison, the proportion of repetition in non-stuttering spontaneous speech was 1.43 incidents per minute.

The repeated words were expected to be shorter than the first-articulated ones (Shriberg, 2001). However, our data did not confirm this assumption. There was almost no difference between the durations of the first and second words in sentence repetition and in narrative. The duration of the second word was about 20 ms longer in conversation (Table 1). There was a remarkable difference between the durations of the first and second words in recall, where the second words turned out to be 46 ms longer on average. The durational difference between the first and second words was significant in the case of recall (one-way ANOVA: $F(1, 121) = 6.184, p = .014$).

Table 1. The durations (mean and range) of the repeated words across discourse types.

<table>
<thead>
<tr>
<th>Discourse type</th>
<th>Duration of words (ms)</th>
<th>First articulation</th>
<th>Second articulation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Mean</td>
<td>SD</td>
</tr>
<tr>
<td>Reading</td>
<td></td>
<td>133</td>
<td>57.85</td>
</tr>
<tr>
<td>Sentence repetition</td>
<td></td>
<td>125</td>
<td>95.48</td>
</tr>
<tr>
<td>Recall</td>
<td></td>
<td>247</td>
<td>89.05</td>
</tr>
<tr>
<td>Narrative</td>
<td></td>
<td>240</td>
<td>108.27</td>
</tr>
<tr>
<td>Conversation</td>
<td></td>
<td>224</td>
<td>125.05</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>230</td>
<td>110.89</td>
</tr>
</tbody>
</table>

The difference between the durations of the first and second words in all three spontaneous speech samples was 24 ms on average, and the difference in length between the second and first words amounted to a total of 73 ms (Fig. 3). The average duration of second words in non-stuttering speech (261 ms, see Gyarmathy, 2009) and in our participant’s speech (253 ms) was very similar; however, the duration of the first words was much greater. The mean duration of the first words was shorter in our PWS’s articulation (230 ms) as opposed to the mean value found in non-stuttered speech (289 ms).
Figure 3.3. Durations of first (left) and second words (right) in repetitions across discourse types (median and range). (R = reading, SP = sentence repetition, RC = recall, N = narrative, C = conversation.)

3.4 Part-word repetitions

The category of fluency failures labeled advancing by Howell (2007) includes repetitions in which only the first part of a word is reproduced. Howell suggests that it is typically content words that are affected in such repetitions, since function words are generally easier to articulate. In our speech samples, content words accounted for 91.4% of all part-word repetitions in sentence repetition, for 78.6% in reading, for 72.3% in narrative, for 63.8% in recall, and for 48% in conversation. The relatively low proportion of part-word repetition of content words in conversation may be due to the fact that in this discourse type, the speaker can plan their words freely and has more time to do so (in contrast to narrative), while others take a turn. Fewer part-word repetitions of content words occur in NFSs’ speech: our PWS’s percentage in the spontaneous speech samples was 61.36%, while the corresponding value in non-stuttered speech was 40.21% (Gyarmathy, 2009).

Across all discourse types, our PWS frequently checked her articulation after producing one or two speech sounds. The proportion of those cases when the speaker interrupted her articulation after one syllable did not show much difference across discourse types, while interruptions after two syllables occurred relatively frequently in conversation (22.7%). Altogether, there were few examples for interruption after three and four syllables in recall and in reading (0.7% and 1.6%).

On average, 8.14 part-word repetitions per minute occurred in our PWS’s speech samples. Her spontaneous speech samples contained 6.04 disfluencies of this type per minute, while sentence repetition contained 11.4 per minute and reading 11.19 per minute. Part-word repetition was much less frequent in recall (7.89 per minute), in narrative (4.04 per minute) and in conversation (6.2 per minute). The frequency of this disfluency in NFSs’ spontaneous speech was reported to be 0.32 per minute (Gyarmathy, 2009).

3.5 Analysis of cutoff-to-repair intervals

The duration of cutoff-to-repair intervals in word repetitions showed large differences mainly in the three types of spontaneous speech samples (in narrative: 117 ms, SD = 129.95, in conversation: 210 ms, SD = 244.04, in recall: 214 ms, SD = 180.52, in reading: 116 ms, SD = 95.92, in sentence repetition: 129 ms, SD = 77.01). The longest cutoff-to-repair intervals were found in recall and conversation, while
the same durations in reading, in sentence repetitions and in narrative were about 100 ms shorter than those in the two former discourse types (Figure 4). Statistical analysis revealed significant differences based on discourse type considering only the spontaneous speech samples (Chi-Square = 11.791, p = .003). Our data confirm that the durations of the cutoff-to-repair intervals of spontaneous speech samples are shorter than those reported for non-stuttering speech (Gyarmathy, 2009). The mean value of the cutoff-to-repair intervals of our PWS was 180 ms, while the mean value for NFSs’ speech was 277.4 ms.

Figure 3.4. Durations of cutoff-to-repair intervals between the repeated words (median and range). (R = reading, SP = sentence repetition, RC = recall, N = narrative, C = conversation.)

The average value of durations of the cutoff-to-repair intervals in part-word repetitions in our PWS’s speech samples (considering all discourse types) was 170 ms, while the mean value of her spontaneous speech samples was 192 ms (in conversation: 130 ms, SD = 148.81, in recall: 157 ms, SD = 272.33, in narrative: 288 ms, SD = 381.69, in sentence repetition: 139 ms, SD = 163.72, and in reading: 185 ms, SD = 220.31), see Figure 5. The longest cutoff-to-repair intervals were found in narrative, with a mean value of 288 ms. The same duration was about 100 ms shorter in reading (with a mean value of 185 ms). None of the remaining discourse types showed any large differences. Since the data were not normally distributed, the Kruskal-Wallis test was used, which confirmed the significant differences based on discourse type (Chi-Square: = 11.431, p = .002). The mean value of the cutoff-to-repair intervals of our PWS was 170 ms, while the mean value for non-stuttering speech was 152 ms (Gyarmathy, 2009).

The disfluent speech flow of a PWS is the end product of several processes that interact in a complex way (Healey et al., 2004). Therefore all measured temporal data of the discourse types were used in a hierarchical cluster analysis (see also Schwartz and Conture, 1988). The goal of this analysis was to highlight the possible connection of discourse types as the result in terms of temporal data in the case of our PWS. Hierarchical cluster analysis was carried out taking into consideration all the analyzed parameters (see the similar application of the analysis in Gahl, 2008). The results show (Figure 6) that reading, sentence repetition and narrative form one
category. Recall and conversation form another, although the connection between them is not as close as that between the discourse types in the former category. Recall shows the greatest distance from other discourse types: this is almost an independent category.

![Figure 3.5. Durations of cutoff-to-repair intervals in part-word repetitions across discourse types (median and range). (R = reading, SP = sentence repetition, RC = recall, N = narrative, C = conversation.)](image)

**Figure 3.5.** Durations of cutoff-to-repair intervals in part-word repetitions across discourse types (median and range). (R = reading, SP = sentence repetition, RC = recall, N = narrative, C = conversation.)

Dendrogram using Average Linkage (Between Groups)

<table>
<thead>
<tr>
<th>label</th>
<th>CASE</th>
<th>0</th>
<th>5</th>
<th>10</th>
<th>15</th>
<th>20</th>
<th>25</th>
</tr>
</thead>
<tbody>
<tr>
<td>reading</td>
<td></td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>narrative</td>
<td></td>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>sentence repetition</td>
<td></td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>recall</td>
<td></td>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>conversation</td>
<td></td>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Figure 3.6.** Results of hierarchical cluster analysis across discourse types considering all analyzed parameters.

The main difference between reading, sentence repetition and narrative as opposed to conversation and recall seems to lie in the different speech planning levels involved in the various discourse types.

### 4 Discussion

The findings of this study supported the assumption that both the incidence and the timing values of the analyzed parameters were indeed dependent on discourse type. It is worth mentioning that the aim of this study was not to analyze the stuttering level of our PWS. We investigated four types of disfluencies that were assumed to be influenced by the analyzed discourse types. There were large differences in the disfluency ratio in our PWS’s speech depending on discourse type.
The least amount of time spent on disfluencies was found in reading (9.8%), followed by sentence repetition (14.2%). Out of the three spontaneous discourse types, disfluencies took up the most time in recall (28.6%); less time was spent producing disfluencies in narrative (21.8%) and the least in conversation (19.2%). However, the occurrence of disfluency types seems to be characteristic of discourse type. We will summarize some of the main facts about the occurrence and the durations of disfluencies depending on discourse type. Filled pause was less frequent in reading and sentence repetition, while it was considerably more frequent in conversation and in narrative. Filled pause was used even more frequently by our PWS in recall, suggesting that she often needed extra time for speech planning in this discourse type (Shriberg, 2001; Watanabe et al. 2008). Prolongations were not very frequent in narrative, conversation and reading, but were frequent in recall and in sentence repetition (Kleinow and Smith 2000). Word repetition did not appear very frequently in reading, and appeared hardly at all in sentence repetition. It occurred most frequently in conversation. Part-word repetitions had a slightly different distribution: they were most frequent in reading and in sentence repetition, and less frequent in the three spontaneous discourse types (Maner et al., 2000; Max and Baldwin, 2010). This finding appeared to support the assumption of a monitoring deficit (e.g., Kolk and Postma, 1997; Hartsuiker and Kolk, 2001; Vasic and Wijnen, 2005; Civier et al., 2010) when the PWS controls her covert errors or assumed errors with oversensitivity in those discourse types where she has no possibility to change or avoid certain words. Next to reading and sentence repetition, most of the part-word repetitions of our PWS occurred in recall. Filled pauses were not only most frequent, but also had the longest duration in this discourse type. Prolongations were longer than filled pauses in all the analyzed discourse types, but they were less frequent. Prolongations were longest in conversation, but least frequent. We can conclude that our PWS used primarily filled pause in spontaneous speech in order to overcome her planning difficulties.

Considering all of the data, disfluencies were most frequent and relatively long in recall, reflecting the complex speech planning processes involved in this discourse type. This claim is further supported by second word durations in word repetitions that were significantly longer than first-pronounced words in this discourse type. Our PWS seems to use the longer duration of the repeated words as a time-gaining strategy (Fox Tree, 1995; Gyarmathy, 2009). Looking at all the data in all discourse types, a careful hierarchy of task difficulty can be assumed for our speaker. Occurrence and temporal patterns of the analyzed disfluencies support the assumption that the speech tasks of reading, sentence repetition and narrative do not differ much from each other, although there are various reasons that might result in speaking difficulty. Recall and conversation were similar with respect to disfluencies, but recall showed the highest number and frequency and longest durations of disfluencies in our PWS’s speech production. This conclusion was also supported by the hierarchical cluster analysis (Schwartz and Conture, 1988).

The assumption that the different tasks posed by different discourse types require different amounts of repairing time was confirmed by durational data relating to the cutoff-to-repair intervals of word repetitions and part-word repetitions. Comparing the durations of the cutoff-to-repair intervals between part-word and word repetitions, three large differences were found. The intervals in narrative were much longer in part-word repetitions than in word repetitions. This suggests that in the
narrative task, our PWS needed more time for the control processes involved in part-word repetitions. On the other hand, the cutoff-to-repair intervals of word repetitions were much longer in recall and in conversation than those in part-word repetitions. This suggests that the cutoff-to-repair intervals of word repetitions were used in these discourse types for speech planning rather than for controlling the produced sound sequences.

The subjective opinion expressed by PWSs about the negative impact of stuttering in certain life situations (Craig et al., 2009) has been confirmed by objective data in our single case study. Our findings may provide a better insight into the inner control and detection of planning failures, as well as their attempted repair in stuttering, a topic worth further study. We have to emphasize that our study is an inherently “weaker” research design and limits the ability to generalize findings to the population at large. However, we think that it is worth analyzing the effect of various discourse types on PWS’ speech involving more participants. The present findings are expected to add something to the discourse type effect measurements on fluency level of a PWS using natural, typical disfluency phenomena. Data for more individuals could be used in multidimensional assessment and treatment of stuttering.

5 Acknowledgements
We would like to thank Dr. Pál Heltai and András Beke for their help in preparing this paper and two anonymous referees for their helpful comments on the earlier versions of the text. This work was supported by the OTKA project No. 78315.

References


AUTOMATIC EVALUATION OF DYSARTHRIC SPEECH AND
TELEMEDICAL USE IN THE THERAPY

Elmar Nöth¹, Andreas Maier¹, Arnd Gebhard¹,², Tobias Bocklet¹, Wilfried
Schupp³, Maria Schuster⁴, Tino Haderlein¹
¹University of Erlangen-Nuremberg, Germany, ²Siemens Healthcare,
Forschheim, Germany, ³m&i Fachklinik Herzogenaurach, Germany, ⁴Klinikum
der Universität München, München, Germany
Contact: Elmar Nöth, Lehrstuhl für Mustererkennung, Martensstr. 3, 91058
Erlangen, Germany, noeth@cs.fau.de, +49 9131 852788

Abstract
After a stroke, the speech quality of the patients is often reduced. This is usually
caused by a deficit of the motor abilities of the vocal tract. The result is slurred
speech. In the various patients, however, very different forms can appear. In the
course of therapy, evaluation of the speech quality is required to determine the
success of the treatment. At the moment, this assessment is performed only
perceptually. This form of assessment is subject to strong intra- and inter-individual
variation. Therefore, an “objective” assessment is not guaranteed. In this study, we
present a rater-independent method for evaluating speech disorders in dysarthria. We
use methods of automatic speech recognition. The idea is to determine the speech
intelligibility – the main outcome parameter of speech – automatically by an
automatic speech recognizer. A correlation of -0.89 was obtained between the
criterion “intelligibility” and the recognition rate of the automatic system, in a
preliminary study. The second part of this paper deals with an additional problem
with this kind of patient. Very often, the stroke leads to partial facial paresis and
generally to reduced mobility. Therefore, it is desirable that therapy sessions are
performed in a telemedical setup. We report on our work towards such a telemedical
diagnosis and rehabilitation system which will allow sessions with a therapist and –
at the same time – diagnose the patient and track the recovery process. We describe
the equipment (web camera, 3-D camera, stereo microphone, and Internet
connection), the patient’s environment, and the working environment of the therapist.
Depending on the network connection, live images of the patient can be sent to the
therapist at a rate of 20 frames per second (fps) at existing LAN connections or 3 fps
with a DSL 6000 connection. At the patient host, a three-dimensional face model of
the patient performing a facial exercise can be generated and transferred to the
therapist in real-time (LAN) or three times real-time (DSL 6000).

1 Introduction
The quality of the speech in patients after a stroke is often reduced (Urban et al.
2001). A deficit of the motor abilities of the vocal tract is usually the cause for
“slurred” speech. There is a considerable variance in the speech outcome across
different patients. In the course of therapy, evaluation of the speech quality is
required to determine the success of the treatment (Ludlow 1994). There is no
objective, validated, automated procedure for the determination of the speech

intelligibility in patients with dysarthria. The perceptual assessment of the intelligibility by speech therapists is not objective and, therefore, subject to inter- and intra-individual variation. In particular, experience is a crucial factor (Paal et al. 2005). In order to obtain a more reliable assessment, patients are often evaluated by an expert committee or panel. However, this is usually performed only for clinical studies and research, because a lot of time and effort are required. In this paper, we first present the use of an automatic speech recognition system to evaluate the intelligibility. Furthermore, we use automatic prosodic features which are also extracted from the speech signal and compare them with a number of other perceptual criteria. There is some previous work on automatic evaluation of dysarthric speech. Van Nuffelen et al. (2009) describe an automatic evaluation method based on phonemic and phonological features. They reach correlations between their objective and perceptual phoneme intelligibility scores from about 0.8 for the two feature groups alone to 0.94 for a combination of the feature groups. Their system is not yet combined with 3-D information (see below) which is favorable for the detailed description of pathologic speech production nor is it provided via internet for telemedical application. For the German language, Ziegler and Zierdt (2008) developed a telemedical system for the evaluation of intelligibility which demonstrated high reliability. They use perceptive evaluation of unknown speech leading to the quantification of speech disorder in accordance to Schiavetti’s claim to quantify the percentage of intelligible words of a word sequence (Schiavetti 1992). However, this telemedical system is based on manpower and therefore includes a time lag until results are given.

The second part of the paper deals with our current work to integrate these diagnosis tools into a telemedical diagnosis and rehabilitation system where the patient can perform a therapy session at home. This is especially important, if the patient has reduced mobility. Apart from analyzing the speech of the patient, we have to provide a fast real-time transmission of the speech signal to the therapist in such a setup. In order to be able to better evaluate the facial expressions of the patient (especially asymmetries due to facial paresis), we create a 3-D model of the patient’s head so that the therapist can have a side view of the patient doing facial exercises. In this paper we concentrate on the transmission speed of our system. The rest of the paper is organized as follows: In Chapter 2, we describe the method for the evaluation of dysarthric speech. In Chapter 3, we describe the patients that we used for the acoustic analysis in this pilot study. The results are presented in Chapter 4. In Chapter 5, we deal with the extension of the system to a telemedical system. In Chapter 5.1., we describe the technology to acquire 3-D information. In Chapter 5.2., the patient’s work place is looked at in more detail, and in Chapter 5.3., results concerning the real-time properties with different Internet connections are presented. The paper ends with an outlook and summary.

2 Evaluation of Dysarthric Speech

The speech data were recorded over the Internet with our “Program for Evaluation and Analysis of All Kinds of Speech Disorders” (PEAKS, Maier et al. 2009). PEAKS runs in any Internet browser and is based on Java technology which allows platform-independent use. The data are transmitted to our server and evaluated centrally (cf. Figure 1). PEAKS currently can only be used as an offline system and is currently not intended for telemedical use, i.e., the patient and the therapist are in
the same location when the therapist accesses the PEAKS web page. Then a client is downloaded which displays the text to be spoken. The patient records his or her speech data which are locally stored by the client. After the recording, the data are securely transmitted to the server. There they are analyzed by a speech recognition system which is based on Hidden Markov Models (HMMs) and a prosody module. As training data for the speech recognition system, solely normal speakers were used. We did not include dysarthric training data in order to be able to better judge the deviation from “normal” speech. The training data were from the Verbmobil project (Wahlster 2000) and covered most regions of dialect of Germany.

The recognizer is described detail in Stemmer (2005). It was developed at the Pattern Recognition Lab (Lehrstuhl für Mustererkennung) of the University of Erlangen-Nuremberg. As features we use 11 Mel-frequency cepstrum coefficients (MFCCs) and the energy of the signal plus their first-order derivatives. The short-time analysis applies a Hamming window with a length of 16 ms, the frame rate is 10 ms. The 12 delta coefficients are computed over a context of two time frames to the left and the right side (56 ms in total). The recognition is performed with semi-continuous HMMs. The codebook contains 500 full-covariance Gaussian densities which are shared by all HMM states. We only used a unigram language model to weigh the outcome of each word model in order to put more weight on the recognition of acoustic features.

The result of the analysis is the number of correctly recognized words with respect to the reference (word correctness WC).

\[
WC = \frac{C}{R} \times 100\% \tag{1}
\]

\(C\) denotes the number of correctly recognized words, and \(R\) is the number of words in the reference.

Furthermore, automatic prosodic features which model energy, fundamental frequency, length of voiced and voiceless segments, jitter, and shimmer were investigated. The prosody module is described in detail in Zeißler et al. (2006). In our case, it takes the forced time alignment of the text to be read (not the recognized text) and the speech signal as input. Thus, the timing information and information about the underlying phoneme classes (such as long vowel) can be used by the prosody module. First, the prosody module extracts the so-called basic features from the speech signal with a frame rate of 10 ms. These are the energy, the fundamental frequency (F0), and the location of voiced and unvoiced segments in the signal. In a second step, the actual prosodic features are computed to model the prosodic properties of the speech signal. For this purpose, a fixed reference point has to be chosen for the computation of the prosodic features. We decided in favor of the end of a word because the word is a well-defined unit in word recognition. The end of a word can be provided by any standard word recognizer, and therefore this point can more easily be defined than, for example, the middle of the syllable nucleus in word accent position. For each reference point, we extract 21 prosodic features. These features model F0, energy, and duration, e.g. the maximal F0 in the current word. In addition, 16 global prosodic features for the whole utterance are calculated. They cover the mean and standard deviation for jitter and shimmer and information on voiced and unvoiced sections. The last global feature is the standard deviation of the fundamental frequency F0. In order to evaluate pathologic speech on a test level, we
calculate the average, the maximum, the minimum, and the variance of the 37 turn- and word-based features for the whole text to be read. Thus, we get 148 features for the whole text. A more detailed description of the automatic speech evaluation system is given in Maier et al. (2009). As the speech evaluation system is more or less unchanged, we would like to focus on the novel combination of the 3-D camera system with the speech evaluation system (see below).

The results of the speech and prosody analysis are available shortly after recording. To compare the human and the automatic evaluation, four speech therapists with at least five years of experience rated the criteria “intelligibility”, “roughness”, and “prosody”. The ratings were performed using a five-point scale assessment for each criterion, and the average per patient was computed in order to obtain rater-independent scores. The agreement between the human and the automatic evaluation was determined as a Pearson (1896) correlation.

Figure 1. Diagram of the client-server architecture of PEAKS (Maier et al. 2009)

3 Patients

For this study, 28 patients with dysarthria were recorded during post-stroke rehabilitation. The patients were 39 to 76 years old. Depending on the severity of the dysarthria, the treatment can take 3 to 18 weeks, with an average duration of 5 weeks. Written informed consent was obtained from all patients participating in the study prior to the examination. Approval was received by the ethical standards committee on human experiments using human subjects at the University Clinic Erlangen.

The data consisted of reading a standard text, the German version (http://de.wikipedia.org/wiki/Die_Sonne_und_der_Wind) of “The North Wind and the Sun” (http://en.wikipedia.org/wiki/The_North_Wind_and_the_Sun). Overall, the text contains 108 words, of which 71 are disjoint. It is widely used in speech therapy in Germany. Figure 2 shows the recording setup using a standard PC at the m&i Fachklinik Herzogenaurach. The audio data were collected using lapel microphones. One is attached to the clothes of the therapist (on the right) and one was attached to the patient’s clothes (on the left). This procedure allows segmenting the speech of both easily. All analyses were performed on the patient’s audio data only.
4 Automatic Evaluation of Dysarthric Speech

The perceptual evaluation of the human raters proved to be very consistent. The inter-rater correlation, i.e. the correlation of one rater and the average of the other three raters was in the range of 0.75 and 0.80. The average of all four raters was 0.78. The results of the evaluation of one rater compared to the average score of the other three raters is depicted in Figure 3. Small amounts of uniform noise are added for better visualization purposes.

The mean of all four experts was used as the reference to train an automatic system. As depicted in Figure 4, there was a significant correlation of $r = -0.84$ between the perceptual assessment of the intelligibility of the four human raters and the word correctness of the automatic speech recognition system ($p < 0.01$). The higher the word correctness, the smaller the human score should be, because ‘1’ on the assessment scale means ‘very good’, and ‘5’ means ‘very bad’. Therefore, the negative correlation is expected. The high correlation is in line with previous studies (Maier et al. 2009).

The evaluations of the criterion “prosody” correlated with the ratio of the length of voiced and voiceless segments, $r = 0.82$ ($p < 0.01$). “Roughness” and the average number of voiceless segments correlated with $r = 0.81$ ($p < 0.01$). The correlation with “jitter” was only $r = 0.66$ (the computation of jitter is described in Levit et al. 2001). Also on the criteria “prosody” and “roughness”, high, significant correlations between human evaluation and automatic prosodic features were found. The relationship of “prosody” and the ratio of the length of voiced and voiceless segments can be explained by the fact that both features are related to accentuation in speech. The correlation between “roughness” and the average number of voiceless segments is also plausible: High roughness may disturb the automatic fundamental frequency extraction algorithm, resulting in an erroneous classification of voiced
signals as voiceless. Hence, a long voiced segment may be divided into several short voiced and voiceless segments. Eventually, this increases the number of voiced and voiceless segments. This hypothesis is supported by the observation that the number of voiced segments also correlates at $r = 0.80$ ($p < 0.01$) with “roughness”.

Figure 3. Correlation of one of the four experts and the average of the other three raters

Figure 4. The correlation between the human experts and the automatic system is high ($r = -0.84$) and significant ($p < 0.001$)
5 Towards a Telemedical System

We have shown that an automatic diagnosis of the intelligibility can be done and this information provides a second opinion to the speech therapist. However, so far, the therapist still has to be present in person. In many cases, this poses a problem since the patients are often immobile because of a stroke that caused the dysarthria in the first place. Thus, it might be very difficult for the patient to come to the therapist’s office. An alternative is that the therapist comes to the patient. This, however, is very costly because a highly trained specialist spends a significant amount of time on the road, especially in rural areas. This is why we have been working on a telemedical solution. Telemedicine is a rapidly developing application of clinical medicine where medical information is transferred through interactive audiovisual media for the purpose of consulting, and sometimes remote medical procedures or examinations. Apart from the obvious ease for the patient and the reduction of costs, a telemedical therapy might actually show better results than a face-to-face therapy. The reason is that the patient is aware of the fact that he has lost fundamental capabilities by only being able to produce slurred speech or not being able to close both eye lids. It might well be that the patient is more uninhibited to show his or her “weakness” to a therapist who sits 300 km away than to perform the exercise face-to-face with a therapist (whom he might even know personally).

In our scenario, we want the speech therapist to work in a hospital and the patient to be temporarily provided with a laptop with audiovisual equipment and fast Internet access. Since the stroke can lead to facial paresis, the visual impression is very important. Dysarthria is a complex and varying disease characterized by a lack of motor control for orofacial, sometimes also respiratory and laryngeal movements, which needs adapted therapy concepts according to the etiology and phenomenology. Quite often other motoric restrictions are combined and locomotion is limited. That’s where telemedical therapy will fit perfectly: to provide medical support to dysarthric patients who cannot visit a speech therapist. The therapist then hears and sees the patient’s articulation and movements and can teach the patient via internet. Of course, therapy then is limited to verbal teaching without thermic or tactile stimulation except for including a naive “co-therapist” who is guided via internet. However, in order to better judge how well a patient can perform an exercise like pursing the lips, the therapist should – at least in some parts of the therapy session – have a 3-D view of the patient. This is why we work on a telemedical system with capabilities far beyond a videophone system. The telemedical system should provide a real-time audiovisual communication of the patient with the therapist, if necessary, provide a 3-D view of the patient’s face, and monitor the progress of the patient with respect to the intelligibility of his or her speech and the quality and symmetry of facial movements. In this way the system can provide help and guidance for computer-assisted practice between the therapy sessions in addition to providing a platform and second opinion for the telemedical therapy sessions. Basis of the new system is our PEAKS client-server platform, which has to be extended with respect to multimodality and real-time capabilities. Figure 5 shows the architecture of the intended telemedical system. The client system for the patient consists of a PC or laptop, Internet connection, stereo microphones, a webcam, a 3-D camera and an illumination source to better control the visual scene. The therapist has the same equipment except for the 3-D camera. In
the next chapter, we have a look at the 3-D camera which is based on the time-of-flight principle.

5.1 Time-of-Flight Imaging

Time-of-flight (TOF) imaging is an emerging technology that provides a direct way to acquire 3-D surface information with a single sensor. Active light sources attached to the camera emit an incoherent cosine-modulated optical signal in the non-visible spectrum of the infrared range (850 nm). The light is reflected by the scene and enters the monocular camera where each TOF sensor element performs a correlation of the local optical signal with the electrical reference of the emitted signal (Xu et al. 1998). Based on this correlation, the phase shift \( \Phi \) representing the propagation delay between both signals is measured. The distance \( d \) can then be computed straightforward,

\[
    d = \frac{c}{2f_{\text{mod}}} \cdot \frac{\Phi}{2\pi}
\]  

where \( f_{\text{mod}} \) denotes the modulation frequency, \( c \) the speed of light. For reasons of periodicity of the cosine-shaped modulation signal, the validity of this equation is limited to distances smaller than \( c/(2 \cdot f_{\text{mod}}) \). At a typical modulation frequency of 20 MHz, the non-ambiguity range is about 7.5 m. The TOF imaging technology benefits from several advantages over other 3-D surface acquisition techniques. The device is compact, portable and easy to integrate. It also provides precise metric information in the sensor coordinate system in real-time, and no calibration steps are necessary. With respect to potential applications in the security, automotive and consumer electronics industry (Kolb et al. 2009), a decrease of manufacturing costs can be expected with mass production being an all-solid-state off-the-shelf technology. Figure 6 shows the principle of a TOF camera.
5.2 A Detailed Look at the Patient’s Work Place

Figure 7 shows a patient during a session. The light is directed towards the patient’s face to have controlled illumination. The speech signal is recorded via 2 stereo microphones at 22.05 kHz (16 bit). The webcam records a 640x480 pixel color image stream (24 bit per pixel). The TOF camera produces a 176x144 pixel depth map (16 bit depth information, 8 bit intensity). The patient either sees the face of the therapist (whenever the therapist wants to demonstrate an exercise) or a control image of himself. The patient can check whether the illumination, the position, and the distance to the recording devices are in a valid range, because this is indicated with an ellipsis around the face. For this the localization of the face both in the 2-D webcam image and the 3-D TOF image is necessary. We use the Viola-Jones Algorithm (Viola & Jones 2001). Normally, only the webcam image stream and the speech signal are transmitted to the therapist in real-time. The number of pictures per second depends on the Internet connection (see below). For certain exercises, e.g., pursing the lips, a 3-D model of the face is created. For this, the TOF depth map and the webcam are registered. After the exercise, this 3-D model is transferred to the therapist. For an exercise lasting about 3 seconds, ca. 480 kB of data have to be transferred. Figure 8 shows a test person during the exercise “showing the teeth”. Here, the viewing direction and the time change from left to right are visible. In the leftmost image, the mouth is still closed; in the central image, the teeth are visible, and in the right image, the mouth is closed again. The viewing direction can be altered by the therapist using the mouse.
Figure 7. The recording station for the patient

Figure 8. Face of a test person during the exercise “showing the teeth” from different viewing directions

Figure 9 shows the work place of the therapist. In the normal mode, only the 2-D image stream is shown (left image in Figure 9). During an exercise, a 3-D face model is computed which is transferred to the therapist (right image in Figure 9) who can then look at the exercise from different viewing angles.

Figure 9. Screen of the therapist’s work place
5.3 Results Concerning Transmission Speed

In the last section, we have described some extensions of PEAKS towards a telemedical system. We have created a mobile workplace that can be used from the home of a patient or in a rehabilitation clinic with a remotely located therapist performing a therapy session or checking whether a patient is correctly performing exercises performed between therapy sessions. In this section, we will describe some experiments concerning the transmission speed of such a telemedical system. For this we performed some test therapy sessions where the test person first read “The North Wind and the Sun” and then performed 4 typical facial exercises which lasted about three seconds each:

- Raising of the eyebrows
- Closing of the eyes
- Showing the teeth
- Pursing of the lips

Even though we have a module that automatically classifies the asymmetry of the face during these exercises (Gebhard et al. 2001), we will only report here on the transmission speed with different Internet connections. Since we only used volunteers without facial pareses during the development of our system, we tested three different scenarios:

- The work place of the patient and the therapist are in one LAN. Transmission speed is 100 MBit (scenario LAN).
- The work place of the patient and the therapist are connected via Internet with a DSL 6000 connection (scenario DSL).
- The work place of the patient and the therapist are connected via Internet with a UMTS stick for the patient’s computer (scenario UMTS).

Using the parameters described in Chapter 5.2., we measured the number of images per second for reading the text and the transmission time for the 3-D models of the exercises. As can be seen in Table 1, the UMTS scenario is not yet applicable without much stronger compression for telemedical applications. In the DSL and the LAN scenario, the transmission speed is fast enough to perform some tests with real patients.

Table 1. Transmission speed for different Internet connections

<table>
<thead>
<tr>
<th>Scenario</th>
<th>Live images per sec.</th>
<th>Duration of 3-D transmission (in sec.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Scenario LAN</td>
<td>20</td>
<td>1</td>
</tr>
<tr>
<td>Scenario DSL</td>
<td>3</td>
<td>10</td>
</tr>
<tr>
<td>Scenario UMTS</td>
<td>2</td>
<td>24</td>
</tr>
</tbody>
</table>

6 Outlook and Summary

We have presented some preliminary work towards a computer-assisted telemedical diagnosis and therapy system for people with dysarthria. The system is
intended for real-time telemedical therapy sessions or for computer-assisted exercise sessions. In both setups, the transmission speed is of crucial importance and it is desirable that an automatic analysis of the patient’s performance can be done, both with respect to the acoustic signal and the symmetry properties of the face. In the therapy session scenario, the automatic analysis acts as a second, objective opinion, by diagnosing the speech of the patient with respect to properties like “intelligibility”, “prosody”, and “roughness” and assessing the symmetry properties of the face during exercises after facial paresis. The system also automatically documents the progress of the therapy. In the exercise scenario, the system has to “supervise” the patient during the exercises that he performs between two therapy sessions and summarizes these sessions for the therapist. In a pilot study, we have shown that the acoustic analysis is in high agreement with the judgment of a panel of experienced speech therapists. We have also shown that the transmission speed is acceptable for a LAN and a DSL scenario. This could even further be improved using compression of the depth data as proposed in Stürmer et al. (2008). We are currently conducting data collection in a rehabilitation clinic with dysarthric patients. This will allow us to verify these very encouraging results with respect to the speech analysis on a larger group of patients and to reevaluate and improve the classification of facial pareses described in Gebhard et al. 2001. There, only a 2-D image was available and we expect significant improvements with the additional 3-D information. During this study, we want to compare the telemedical setup with a face-to-face setup.

Furthermore, we will develop instruction dialogues so that the system can analyze exercises between therapy sessions. Finally, we want to improve the UMTS scenario, i.e., increase the compression rate until the transmission speed is acceptable and test whether the quality of the images is good enough to allow a telemedical live session.

References


1 Introduction

Speech production begins with the intention to speak and ends with actual pronunciation. In spontaneous speech, the message is not precompiled; the formation of ideas and articulatory implementation occur simultaneously. While the speaker articulates a unit, she also plans the following train of thoughts in terms of content, grammar, phonology, and articulation. This simultaneity of operations may result in difficulties at any level of the process: disharmonic surface phenomena or disfluencies (Gósy 2005). Disfluencies fall into two major groups: uncertainties, and implementation errors.

The linguistic study of disfluency phenomena started in the 8th century A.D. (Al-Ki-sa’i, cf. Berko Gleason & Bernstein Ratner 1998). European linguistics has been dealing with it for over a century now, with research on English and German (cf. Mahl 1956, Goldman-Eisler 1958; Maclay & Osgood 1959; Bernstein 1962; Boomer 1965), and later other languages (cf. Ferber 1993; Nadeau 2001). Research on Hungarian began in earnest in the late 1990s (Gósy 1998; Huszár 1998) and has become an important area, both in psycholinguistics and phonetics, due to the corpora that have been and continue to be compiled.

Disharmonic phenomena provide information about processes and operations that cannot be directly observed by analysing error-free speech material. Disfluencies are “windows” on the hidden processes of speech production (Fromkin 1973), because the same production rules are responsible for both normal and erroneous forms noted in speech (Pouplier & Hardcastle 2005). The study of disfluencies yields important information about overt or especially covert self-monitoring processes and the way errors are self-corrected, as well as information about speech planning processes, the self-monitoring mechanism, corrective processes, etc.

The present study aims to describe self-monitoring and corrective processes in spontaneous speech by investigating two kinds of uncertainty disfluency (repetition, restart) and three kinds of error disfluency (false start, wrong word choice, perseveration). Reflecting self-monitoring processes with the properties of disfluencies provides us with quasi-objective data about how the speech planning process works, thus allowing strong conclusions.

2 Structure of dissertation

The dissertation consists of 15 chapters. It includes analyses of self-monitoring and repair strategies from the perspective of speakers and listeners in typical and non-typical situations, and 5 types of disfluency.

Chapter 1, the Introduction, includes a description of the topic and the general aims of the research. Chapter 2 discusses the connection between speech and the brain, the speech production process and some related speech production models. Chapter 3 introduces the characteristics of spontaneous speech, its types, issues
related to naturalness, and the most important, a Hungarian speech corpora. Chapters 4 and 5 deal with disfluency phenomena. Chapter 4 gives a detailed survey of types of disfluency stemming from a speaker’s uncertainty. Chapter 5 is an overview of the operation of the self-monitoring mechanism, theories about it, and the repair strategies that speakers or listeners employ.

Chapter 6 states the dissertation’s aim and its main hypotheses. The main goal of this study was to explore the correction processes that speakers and listeners employ in communication. I wanted to discover (i) which types of disfluency disturb speech comprehension and which facilitate it; (ii) whether these are in harmony with the speakers’ correction processes; (iii) what changes occur in self-monitoring due to the individual’s current state of mind and/or the acoustic environment; and (iv) what characteristics the repair of individual uncertainties and errors exhibit. My hypothesis is that speakers and listeners only correct some errors, depending on their type. We assumed that altered physiological status (drunkenness) and acoustic environment (noise) would affect self-monitoring and correction processes and that the individual uncertainties and errors would differ with respect to the method and duration of repair.

Chapter 7 introduces the general methodology of the experiments, the subjects involved, and the corpora used. Chapter 8 discusses the results pertaining to the factors affecting disfluency phenomena and self-monitoring, and consists of two major sections dealing with the effects of alcohol and noise. Chapter 9 contains research results about the effect of disfluency phenomena on the speech perception mechanism. Chapter 10 presents results concerning the two types of uncertainty and three types of errors studied in this dissertation, in four major sections presenting an analysis of repetitions and restarts, false starts, cases of wrong word choice, and perseverations. Chapter 11 includes specific and general conclusions. Chapter 12 is a brief summary of the dissertation. Chapter 13 contains the theses of the dissertation. Chapter 14 is a list of references. Chapter 15 is an Appendix containing the series of pictures used for the experiments concerning the effect of noise, and the text of the perception test.

3 Material, methods, and subjects

Approximately 44 hours of recordings from the BEA database involving 144 participants were analysed with the aim of exploring the two uncertainty-based disfluencies, the three error phenomena, and the speakers’ and listeners’ strategies. To study the effect of alcohol and noise on spontaneous speech, we created two new corpora with 29 speakers, totalling 6 hrs 7 minutes. Acoustic phonetic analyses were performed with Praat 5.0.03 (Boersma & Weenink 2009), and statistical analyses were performed with SPSS 13.0.

4 Repair strategies in various speech situations

Spontaneous speech is affected by a number of factors. This study includes a detailed investigation of the effects of a noisy environment and alcohol on spontaneous speech and self-repair.

The effects of four different types of noise on spontaneous speech (20 female subjects; 60 minutes and 49 seconds of recordings) were investigated: dog bark, dental drilling, jackhammer and music. The results showed that the Lombard-effect (the tendency to increase one’s vocal effort in response to noisy conditions; cf. Gósy 2008) asserts itself irrespective of the type of noise involved; but no further clear conclusions can be drawn concerning how noise affects the speech process. At any rate, under noisy circumstances, speech production (and not only
speech comprehension) is more difficult. Noise type and intensity, how it is transmitted (via air vs. headphones), etc. are all relevant.

The relative frequency of errors and uncertainties also changed. Disfluencies of uncertainty occurred more often in noise than in silence, indicating that although speakers had more problems in speech planning, they also controlled their speech more intensively because of the noise and therefore planning disharmonies did not surface as actual errors, as detection and (covert) repair of potential errors are signalled by uncertainties.

In contrast to our preliminary assumptions, less disfluency phenomena occurred per minute in noisy speech than in the silent condition. The occurrence of disfluencies per minute did not increase, but the negative effect of background noise was shown by the fact that all four groups of subjects produced a more varied array of errors in the recording under noise than in silence.

The results revealed that inhibition of acoustic feedback has an adverse effect and this dissertation is the first study of its kind in Hungarian. As expected, the level of adaptation. In cases of prolonged background noise of a constant level diminish (i.e., the participant ‘adapts’ to the noise). I found no significant differences between the noisy and silent portions of speech, which can be accounted for by auditory ‘adaptation’ (Tulipánt 2004). People become accustomed to the background noises accompanying their verbal communication in everyday situations.

Little research has investigated the effect of alcohol on spontaneous speech and self-monitoring (cf. Braun et al. 1992; Braun & Künzel 2003; Stemberger 1993), and this dissertation is the first study of its kind in Hungarian. As expected, the negative effect of alcohol on motor coordination (cf. Környey & Kassai-Farkas 2002; Buda 1998) led to problems in speech coordination, whereby the number of error-type phenomena increased.

![Figure 1](image-url)  

**Figure 1.** Rate of correction of disfluency phenomena in silence vs. under noise of diverse types

The effect of noise depends on the person, the type of noise, its intensity, and the level of adaptation. In cases of prolonged background noise of a constant level of intensity, the listener adjusts to the noise so that the effects of the background noise diminish (i.e., the participant ‘adapts’ to the noise). I found no significant differences between the noisy and silent portions of speech, which can be accounted for by auditory ‘adaptation’ (Tulipánt 2004). People become accustomed to the background noises accompanying their verbal communication in everyday situations.

Little research has investigated the effect of alcohol on spontaneous speech and self-monitoring (cf. Braun et al. 1992; Braun & Künzel 2003; Stemberger 1993), and this dissertation is the first study of its kind in Hungarian. As expected, the negative effect of alcohol on motor coordination (cf. Környey & Kassai-Farkas 2002; Buda 1998) led to problems in speech coordination, whereby the number of error-type phenomena increased.
The most frequently occurring phenomena in both conditions (alcohol and control) were expletive words and prolongations, but the speakers used different strategies for gaining time under alcoholic and control circumstances. While in the control condition, the time needed for self-monitoring and resolving disharmony was provided by hesitations and prolongations. In the alcoholic condition repetitions, restarts, and expletive words fulfilled this function. The two processes appeared to be interconnected: alcohol makes the individual verbose and she or he uses speech to buy time for mental lexicon searches and resolving the inconsistencies arising during linguistic planning. Subjects are less capable of harmonising planning and execution, resulting in both higher frequency and variance of occurrence of disfluencies per minute in the alcohol condition than in the control one. The ratio of expletive words, repetitions, false starts and simple slips of the tongue increased the most in the alcohol condition. Analysis of the errors showed that non-sober speakers had the most difficulty with content planning, the mental lexicon’s proper activation, linguistic planning, and articulatory movement coordination.

One-fifth of the disfluency phenomena of the corpus are correctable disfluencies. Under the influence of alcohol, the frequency of their occurrence increased by 7.79%. The subjects were not able to recognise and correct all their errors, whether they were sober or otherwise (cf. Figure 2).

According to our statistical data, male speakers produced them significantly more to be aware that lexical access is more difficult than while being sober, and they concentrate more on that process: before the whole word is finished, they try to recognise and correct their errors. However, participants under the influence of alcohol seem they otherwise would. However, participants under the influence of alcohol seem to be aware that lexical access is more difficult than while being sober, and they concentrate more on that process: before the whole word is finished, they try to recognise and correct their errors. However, they are unable to correct them in all cases. Thus, self-monitoring strategies are altered under the effect of alcohol.

**5 Listeners’ repair strategies**

In the research on perception-based repair mechanisms, recordings involving 11 speakers were analysed. The frequency of disfluencies varied across speakers. According to our statistical data, male speakers produced them significantly more often than females. The results confirmed that the majority of disfluency phenomena occurring in spontaneous speech (specifically, three out of four) go
unnoticed by listeners. During speech perception, unintentional (covert) repair is mostly successful. In accordance with my hypothesis, what listeners identified as disfluencies were longer silent pauses (around 500 ms), as well as phenomena that were expected to cause problems for perception and comprehension (ordering errors, false starts, cases of wrong word choice). Shorter silent pauses (250 ms or less) were not perceived by the listeners, suggesting that they are there to provide the listener with processing time. This proposal is also supported by the fact that the listeners were only able to identify one-fifth of the physically present silent pauses.

The success of repair processes operating during speech processing depends on the individual listener: one subject was able to identify almost 41% of disfluencies, whereas another one was merely able to spot 7%. The recognition of disfluencies also depends on the type of phenomenon and to some extent on its frequency of occurrence (Figure 3).

![Figure 3. The occurrence of disfluencies in the test material, and the rate of their recognition (the left-hand-side vertical axis shows the number of disfluencies per minute, the right-hand-side axis shows their recognition rate in percent)](image)

The phenomena most often occurring in the test material (repetitions, hesitations and false starts) were perceived in 40–50% of the cases. According to our statistical analysis, the type of disfluency is decisive with respect to its recognition. The listeners unambiguously identified hesitations, false starts, and perseverations, whereas grammatical errors went unnoticed most of the time. For the listeners, disfluency phenomena based on uncertain planning were less conspicuous: they noticed only 35.99% of these, while they identified 57.38% of the speech errors. This can be explained by the communicative function of the various phenomena: it is mainly those belonging to the category of ‘disfluency due to uncertainty’ that provide the necessary processing time for the listener. Of the errors, it was false starts, wrong word choice cases, perseverations, and instances of the tip of the tongue (TOT) phenomenon that disturbed perception the most. These errors, due to their sense-distorting nature, are harder for listeners to repair automatically.

Ungrammatical forms caused the fewest problems in perception. Perception reacts to word and phoneme level errors, as well as those concerning articulatory planning (segmental order), far more than to grammatical incoherence. I can
conclude that speech processing relies on the harmony between sound shapes representing root morphemes and meaning, that is, on lemmas, rather than on grammatical structure.

Comparative analysis of the subjects’ production and perception shows that the two processes are not mutually independent: listeners have more success in identifying disfluencies that are less characteristic of their own speech, but the subjects exhibit large individual differences in their perception processes.

6 The success of self-monitoring and disfluency types

To determine whether the two main categories of disfluency phenomena also differ with respect to self-monitoring processes, the dissertation contains a detailed analysis of two uncertainty-based and three error-type phenomena. For all five types of disfluency, I analysed the length and structure of the editing phase (i.e., the time that correction takes) in terms of variables like lexeme type (content vs. function word), part-of-speech, the connection between the error and its repair, etc. These results supported the claim that the length of correction time depends on the type of disfluency (Figure 4); a conclusion that is also supported by statistical investigations (one-way ANOVA: F(4, 1044) = 28,881; p < 0.001).

It is not always easy to know if repairs result from covert or overt self-monitoring. However, when no editing phase was detectable, i.e., when the lag was 0 ms long, it could be assumed that the correction was part of covert monitoring. The types of disfluency I analysed did not differ significantly in this respect. For all types, the speakers repaired 30–40% of the phenomena studied in covert self-monitoring. On the other hand, from the length of editing phases, we can conclude that the self-monitoring mechanism works considerably worse in the case of errors resulting in repetition and in cases of wrong word choice. Thus, most errors can only be corrected on the basis of speakers hearing their own speech.

Detailed acoustic and temporal analyses have furthermore indicated that it cannot be unambiguously decided whether a given phenomenon reflects the speaker’s uncertainty or if it is a specific error type. Restarts are conventionally classified as cases of uncertainty; but our data suggest that restarts involving content words in fact go back to problems of lexical access. In these cases, as in cases of errors in the activation of the mental lexicon, the monitor gives an error signal during articulation, which therefore stops the process. If the self-monitoring mechanism then classifies the given word as error-free, the phenomenon surfaces as a restart; whereas if it is found to be erroneous, there is a case of false start.

In all types of disfluency, the structure of editing phases determines their duration. If more time is needed for error correction, the speaker may use time-gaining strategies. On the other hand, the editing phase was shorter when function words were involved, confirming the claim that they are stored in the mental lexicon as clichés, hence both their access and repair is faster than those of content words.
Figure 4. The length of the editing phase with the disfluency phenomena analysed

7 Conclusions

The purpose of this dissertation was to explore the correction processes made by speakers and listeners under different circumstances (in noise, under alcohol and control circumstances). I analysed a large amount of spontaneous speech, and the relevance of the data was confirmed by statistical analyses. On the basis of the results, the following conclusions can be drawn.

1. Speakers correct only some of their disfluencies (50–60%), and correction varies with the type of disfluency. Speakers focus more on correcting errors that occur early in the planning process (false start, wrong word choice); grammatical errors and articulatory slips are only rarely repaired.

2. Factors affecting spontaneous speech in general exert an influence on the self-monitoring mechanism, too. The disfluency correction process is adversely affected by some features of the communicative situation (noisy environment, effect of alcohol).

3. Phonetic and temporal analysis of the five types of disfluency phenomena indicates that the duration of repair (the length of the editing phase), as well as the realisation of covert vs. overt monitoring depends on (i) the type of disfluency, and (ii) whether the given phenomenon concerns function or content words.

4. Listeners recognise only a small fraction of disfluency phenomena in spontaneous speech (one out of four); hence, disfluencies in general do not result in processing difficulties in speech comprehension.

5. The processing of the various disfluency phenomena in speech perception is not uniform. Uncertainty-based phenomena are less conspicuous, because their function is mainly to provide the listener with sufficient time for processing what has been said. Error type phenomena, on the other hand, are more distracting for the listener. The recognition of the specific disfluency phenomena depends on the type of disfluency concerned. Listeners tend to identify false starts, wrong word choices and perseverations more readily than other types of communication issues because disfluencies of the types noted above lead to a sense distortion in communication, whereas grammatical errors are hardly noted.

References

INTERNATIONAL COLLABORATION BETWEEN THE DEPARTMENT OF PHONETICS/PHONETICS INSTITUTE OF ELTE AND THE BILINGUAL AND CROSS-LINGUISTIC LANGUAGE LABORATORY AT THE UNIVERSITY OF HOUSTON

Conferences are more than venues where professionals in the field congregate to present their research; they are also places where people meet to network and forge new research alliances. The 13th Meeting of the International Clinical Phonetics and Linguistics Association in Oslo in 2010 brought about the beginning of such a research collaboration between Judit Bóna and Mária Gósy of the Department of the Eötvös Loránd University and Ferenc Bunta of the Department of Communication Sciences and Disorders at the University of Houston. All of us have been working on various aspects of phonological and phonetic development and have read each other’s works with great interest, but only in the summer of 2010 did we begin discussions to join our efforts and skills to tackle important problems in monolingual and bilingual English and Hungarian phonological development.

Individually, each of us is working on somewhat different aspects of language acquisition. Dr. Bóna, a former student of Dr. Gósy, has been working on age characteristics of speech production and perception and temporal aspects of speech. Dr. Gósy has been working on L1 acquisition, speech perception, phonetics and psycholinguistics of speech production. Dr. Bunta has been investigating various aspects of bilingual phonological development in Spanish-English and Hungarian-English bilinguals in the US. Heeding Aristotle’s advice that “The whole is greater than the sum of its parts”, we are joining forces to expand our research into areas that remain unexplored or underexplored, such as phonological acquisition in bilingual and monolingual Hungarian-English speaking children. To date, our collaboration has born joint participation at the Speech Science Conference in Budapest, Hungary in the fall of 2011, and we are currently working on multiple co-authored works, including an article on the speech production skills of monolingual Hungarian children with speech sound disorders and a chapter for the new LARSP book (“Profiling Grammar: More Languages of LARSP”).

The future of our collaboration is very bright not only because of the novel aspects of investigating phonological development in Hungarian-English bilinguals as compared to their monolingual peers, but also due to the fact that Hungarian and English have markedly different phonologies. The findings of our on-going and planned studies have potentially far-reaching ramifications for speech science, phonological acquisition theories, and for clinical practice in speech-language pathology.

Ferenc Bunta
Department of Communication Sciences and Disorders,
The University of Houston
e-mail: fbunta@Central.UH.EDU
The International Congress of Phonetic Sciences is organized every fourth year in a different country. In 2012, the conference took place in the wonderful town of Hong Kong, China. This was the first time that the conference was hosted by an Asian country in the history of this event. Given the large number of submitted papers, a significant number of papers were accepted: 7 plenary, 31 special, 311 regular oral, and 231 poster presentations. This is the first time that the conference proceedings has been made available on the internet: (http://www.icphs2011.hk/ICPHS_CongressProceedings.htm).

The plenary lectures covered a variety of phonetic topics. Professor Klaus Kohler talked about the intertwining of segmental and suprasegmental features with respect to communication functions. Professor Sara Hawkins introduced new results on using phonetic details of lower advantage if they are task-related. These results support speculations about modeling speech perception as part of a biologically-grounded, situated theory of human interactive behavior. Professor Jan Maddieson gave a presentation on the complexity of phonological systems and subsystems. He suggested that various subsystems show a great degree complexity. In addition, he introduced how compensation might be found based on relative frequencies. He recommended that the global distribution of phonological complexity might deserve further exploration, which would include investigating environmental and demographic factors. Professor Daniel Recasens talked about the need for collecting extensive experimental data on the language dependent articulatory and acoustic characteristics of vowels and consonants. He showed the relevance of this question with data on allophonic realizations of the same speech sound in several language groups. Professor Jacqueline Vaissière discussed considerations of the vowel space, and the cardinal vowels in both acoustics, and articulation, and the relationships among these aspects. Professor Randy L. Diehl gave his talk on the robustness of speech perception. He and his colleagues found that the phonetic properties of sound categories covary since language communities tend to select properties that have mutually reinforcing auditory effects (auditory enhancement hypothesis), thereby creating intermediate perceptual properties that serve, individually and in combination, as the basis for distinctive features. Based on their data, human perception may be hypothesized to be optimized with respect to the distributional properties of naturally produced speech sounds. These two facts support the idea of perceptual robustness. Professor Louis C.W. Pols’ topic covered speech dynamics. Acoustic and perceptual tests had been carried out, and the results were discussed in terms of effecting factors, and consequences of such for both speech recognition and speech synthesis.

The special sessions also involved a variety of diverse phonetic questions. The first special session was titled ‘Expanding Phonological Horizons: On the Role of Aerodynamics in Phonology’. Aerodynamic constraints and their role in phonological systems, and their relevance were demonstrated, as well as, analyzing voicing constraints and voicing initiation gestures, and the velic opening gestures.

The second special session (‘Phonetic Fieldwork’) served to consider the role and future possibilities of field work in phonetics. The presenters talked about specific problems like perception, aerodynamic measurements, and laryngoscopic analysis in field work, and also considered challenges related to eliciting data and ethical considerations in the Pacific Northwest.
The third one (‘Phonetic Teaching and Learning: Recent Trends, New Direction’) dealt with phonetic teaching and learning. The specific topics presented were the possibilities of new techniques, surfaces (e.g. YouTube), the new trends and directions, and also an overview of the specificity of teaching English as a Lingua Franca, actualized for East Asia.

The special session of ‘Shapes and Tones – Towards a More Holistic Perspective in Intonation Research’ raised questions about both production and perceptual points of the field. One of these was how listeners cope with an interrupted pitch contour, i.e. how they perceive e.g. an unvoiced segments *contour*. The results suggested that these breaks are simply ignored as opposed to being interpolated or extrapolated. Also the question of tone realization and its perception and categorization was addressed in this session. The categorical perception of Porteño nuclear accents, and the production strategies of pitch accent categories were discussed.

A special session was addressed to ultrasound studies of speech production (‘Ultrasound Studies of Speech Production’). The talks revealed several tricks/techniques, and the newest results on the topic. The gesture timing of vowel-nasal sequence production was shown to be dependent on diverse factors. The results revealed that tongue and syllable position play a role in gestural timing, and while some of the data can be interpreted by biomechanical models, others seem to be perceptually linked. These methodological presentations shed light on data synchronization and analysis.

One of the special sessions (‘Prosodic Focus: Cross-linguistic Distribution and Historical Origin’) dealt with the cross-linguistic distribution and historical origin of prosodic focus. The introduced researchers focused on the phonetic realizations, perceptions, and bilingual effects of focus and post-focus positions. The phonetic realizations and their relevance in perception were introduced with Turkish focus types, and with six dialects of Chinese. Another research project found that bilingualism (English-Chinese) does not cause carry over of post-focus compression from one language to other. The last presentation of this session discussed the question of post-focus compressing and non-compressing languages. The current state of the art of these investigations of the topic, the resulting hypotheses – especially that of the historical development of the languages – were reviewed in the talk.

Two special sessions (‘The History of Phonetics’ Part 1 and 2) were dedicated to the history of phonetics. The talks reminded us of important past research. We heard about a pioneer of speech synthesis, Christian Gottlieb Kratzenstein and about the possibly first person to investigate VOT\(^1\), Adjarian. The instruments and measurements of our predecessors were introduced and reconsidered from various aspects. Kempelen’s speaking machine was discussed from the point of view of present application possibilities for demonstration and research. While another presentation analyzed what measurement accuracy the early phonetic experiments may have had. The history of Hungarian phonetics was also introduced. Our trip through our predecessors started with Kempelen’s work and followed up with the various fields of phonetics to the present. Another presentation helped us to think over the development of vowel systems. Several concepts were discussed, and compared with regard to acoustic features, perception, and the interrelationships among these aspects. One of the presentations allowed us to travel back approximately 90 years. Some films had been found at UCL. The film introduced at the conference showed an early – possibly prototype – kymograph and

\(^1\) This presentation is not involved in the proceedings, but can be read in the paper: Angelika Braun, "VOT im 19. Jahrhundert oder "Die Wiederkehr des Gleichen"", *Phonetica* vol. 40. 323-327. 1983
sensitive flame in functioning. The organizers of the special session intend to establish a special interest group for research and discussion of the history of phonetics.

Other presentations introduced results on topics like speech production, speech acoustics, speech perception, speech physiology, speech prosody, phonation, voice quality, phonetics of connected speech, phonology, phonetic universals and typology, sociophonetics, sound acquisition, phonetics of L1-L2 interaction, field methods in phonetics, speech synthesis, speech recognition, clinical phonetics, forensic phonetics, neurophonetics, etc.

The next conference will take place in picturesque Glasgow in 2015.

Tekla Etelka Gráczi
Research Institute for Linguistics, Hungarian Academy of Sciences,
Budapest, Hungary
e-mail: graczi.tekla.etelka@nytud.mta.hu
The International Speech Communication Association (ISCA) conference was held in Florence, Italy between 27 and 31 of August, 2011. INTERSPEECH is a great opportunity for speech technology researchers, instructors and students to present their latest results and to discuss their research. The conference is organized by ISCA annually in a different country. Recent locations prior to the 2011 event were: Japan, United Kingdom, Australia, Belgium, USA, Portugal, Korea and Sweden.

The theme of the 2011 conference was “Speech science and technology for real life”. The topics of the plenary talks displayed an emphasis on understanding the structure of conversation, properties of spoken language and their modeling by machine.

INTERSPEECH 2011 has already broken a record: 1439 papers were submitted, making it the largest spoken language conference ever in the long history of these interdisciplinary events. Approximately 57% of the regular papers were selected for presentation as a result of the peer review process using 3 or 4 reviewers for each paper.

The topics of the presentations were extremely diverse, such as speech recognition, speech synthesis, speech perception, speaker recognition, spoken language understanding, speech production (articulatory modeling), physiology and pathology of spoken language, paralinguistic information, prosodic modeling, and so on. Speech technology is increasingly focused on modeling conversation, automatic analysis of dialogues, and speaker diarization. It seems that these issues will be among the most frequently investigated topics for the next few years when it comes to speech communication research.

The research of the Hungarian Research Institute for Linguistics, the Department of Telecommunications and Media Informatics of the Technical University of Budapest and the Institute for Psychology of the Hungarian Academy of Sciences was represented at the conference with two presentations: “Context and speaker dependency in the relation of vowel formants and subglottal resonances – evidence from Hungarian” (by Gráczi, Tekla Etelka, Lulich, Steven M., Csapó, Tamás Gábor and Beke, András). The title of the other presentation was “Analysing the correspondence between automatic prosodic segmentation and syntactic structure” (by Szaszák, György, Nagy, Katalin and Beke, András).

INTERSPEECH 2012 will be held in Portland, followed by Lyon (2013), and Singapore (2014).

András Beke
Research Institute for Linguistics, Hungarian Academy of Sciences, Budapest, Hungary
e-mail: beke.andras@gmail.com
BOOK REVIEWS

Lingua Americana July-December 2010 year XIV, no. 27, and Lingua Americana January-June 2011, year XV, No. 28.
(Published by the Universidad del Zulia, Maracaibo, Venezuela, 117 pp., Index of articles of the 2010 volume (pp.118-122), Instructions for authors in English and Spanish (2+2 pp.), ISSN 1316 6689, no price given).

Reviewed by
Judith Rosenhouse
Swantech Ltd., Haifa, Israel
e-mail: swantech@013.net

This journal (whose editor is Professor Godsuno Chela-Flores, who writes the editorial in each issue of the journal) is a linguistic review of the Institute of Literary and Linguistic Research at Universidad del Zulia, Maracaibo. It deals with three main areas of research: general linguistics, applied linguistics and indigenous languages. Each paper has abstracts in Spanish and English, and one of the papers has also an abstract in French. Most of the papers are written in Spanish, some in English and one in these two issues is in French. Each of these two recent issues of Lingua Americana includes a few papers (5 in the 2010 issue and six in the 2011 issue). We review here only those related to phonetics/phonology, i.e., four papers from the 2010 issue and one from the 2011 issue.

Lingua Americana July-December 2010:

The first paper is “Linguistic variables in anomalous segmentation for Maracibo children” by Eglee Coromoto Gonzalez Urdaneta (pp. 11-31). The study is related to speech in the area of interaction between spoken and written communication. The author studied the writing of 120 elementary-school children by studying several grammatical variables including pronouns, prefixes and suffixes. The data from the children’s written works were collected by the program SHOEBOX and analyzed statistically. The results relate to the issue of word definition and examines three hypotheses: the sequence C#V and V#V would show major “hypo-segmentation”; functional words would present less hypo-segmentation than content words; prefixes would tend to hypo-segmentation less than suffixes. The studied elements included pronouns such as ‘mi, tu, su’, definite and indefinite articles ‘la, los, un,’ prefixes such as super, extra, se, micro’ and words sequences such as ‘se escapan’, ‘por un’, ‘en el.’ The study indeed shows the influence of linguistic variables on the anomalous segmentation on these children’s writing, for which the author suggests some pedagogic teaching strategies which take into consideration the manner of speech.

The following paper is by Bertha Chela-Flores, “Distributed practice and pronunciation teaching: Recycling the instruction within Speaking Oriented Activities” (pp. 35-50) The author criticizes a teaching manner that teaches foreign language words and sentences but does not lead to good pronunciation in spontaneous speech, and describes a method that will teach sentences in real texts while taking in account the speech rhythm of that language as part of its prosodic structure. This rhythmic element needs to be separately taught by first tapping out unstressed and stressed syllables as found in various basic utterances. She demonstrates this method using examples of texts in English (also given in the appendix), including realistic dialogs, with the stressed/unstressed syllables
marked (as in some poetry studies). Teaching such elements should be enhanced by repeated occurrences of the elements in more than one lesson. In the concluding remarks, Chela-Flores remarks that “a proper rhythm also implies the proper blending and linking of words, the pronunciation of final consonants and consonant clusters” – apparently reflecting typical errors of native speakers of non-English languages. She describes this “distributed practice” following a learning-psycho-logical method a non-linear approach to pronunciation teaching and adds that empirical data of this method are under study.

Pue-Fang Fung is the author of “An instrument for the analysis of academic writing passed on the characteristics of primary orality” (pp. 51-74). The starting point of this study is the relation between written communication and speech as examined in written texts by university students. The discourse structure of these texts shows elements that are characteristic of spoken language – more conjunctions and less subordinating particles, i.e. fewer subordinating sentences, or unfinished sentences. The study analyzes academic texts using the nine characteristics of primary oral culture (Ong, 2011) and finds 22 descriptors of such characteristics, each of which is described and exemplified. The author discusses discourse from pragmatic and theoretical perspectives and considers the method as an analytical instrument which may replace the grammatical-prescriptive approach and explain idiomatic inconsistencies in written discourse.

Inés Blanco Sáa, “The problem of the interprosody of Venezuelan learners of French” (pp. 75-103). This paper, written in French, considers some phonetic differences between Venezuelan speakers’ French (prosodic inter-language) and their native Spanish. The author first presented a contrastive analysis of French and Spanish prosodic features in sentences with different emotional and pragmatic features (speech acts, in her phrasing). Based on audio analysis of a non-spontaneous Spanish speech corpus and perception tests, the French speakers recognized various sentences (mainly questions, by their final rise, but the results were not so good in polite questions or wh- questions). Forty-five native French speakers were examined as to French spoken by Venezuelan speakers in the second test: They had to answer whether the speaker they heard in several recorded sentences was a French native or foreign speaker. Then an experiment was carried out by recording sentences from eight Venezuelans speaking French, eight Venezuelans speaking Spanish and eight French speakers speaking French. The recordings were inspected for duration, F0 and intensity using WinPitch V.1.8 and Speech Analyzer. Several examples of questions are shown in the paper as differences between the groups. Differences were found between the Venezuelans and French speakers’ questions by higher sentence-final F0, but these are not borrowed from Spanish. The author concludes that this is probably due to perceptual errors. Other differences are due to inter-language differences of F0 and durations. Sáa remarks in the end that it would be desirable for foreign language learners to acquire both an “efficient” and “agreeable” speech manner, in which the role of prosodic features is uncontestable.

In Lingua Americana January-June 2011 we find:

Migdalía Durán, in “Polysystemic interpretation of an acoustic characterization of the postvocalic alveolar fricative /s/ in the speech of young people in Barquisimeto” (pp. 56-80), analyzes a phonetic topic, namely, the pronunciation of the alveolar fricative /s/ in a certain urban location in Venezuela. (This phoneme has been weakened down to the glottal /h/ or total deleted in several Spanish dialects of Venezuela and other countries in South America, Middle America and Mexico.) Durán’s study is conducted using an acoustic analysis
method (following Llisterri, 1991), which also considers sociolinguistic factors. In the beginning, the author describes the “natural poly-systemic” model (Godsuno Chela-Flores, 1983, 1987) which predicts various phonological changes, specifically depending on inter-syllabic, post-nuclear, post-vocalic contexts. The research material was based on the formal and informal speech recordings of two 24 year old (male and female) speakers from Barquisimeto where this feature in the local dialect has not been studied. Sentences and words with /s/ were found in the recordings and analyzed acoustically (using Speech Analyzer version 2.5) and then statistically (using SPSS). The author describes three varieties of pronunciation of this phoneme: strong glottal /h/, weak glottal [ʰ] and zero, i.e., total elision of /s/. For example, the frequency of use of the strong and weak glottal and elided /s/ in 308 analyzed cases is: strong 16.2%, weak 45.9%, and totally elided 38% (p. 69). Differences were found between the male and female speakers, in postvocalic final and medial context of the implosive /s/. Formal and informal style differences were also found in the pronunciation of /s/ as /h/ or zero. Several other differences relate to other features of the pronunciation of this phoneme. The paper ends with a comparison of these findings with works on other dialects in South America. Durán notes that not enough research has been done on the dialects of Venezuela, not even on the pronunciation of /s/, although some works are mentioned in the paper. The author calls for more work on the varieties in Venezuela, which will also support phonological theory development.

In sum, these two issues of Lingua Americana have provided some interesting perspectives on issues which relate to general phonetic issues, not limited only to South America.

References

Kreiman, Jody, and Diana v.L. Sidtis (2011)
Foundations of Voice Studies
Malden, Mass., USA: Wiley-Blackwell, 504 pp. $ 124.95

Reviewed by
Sandra Ferrari Disner
Dept. of Linguistics, University of Southern California

This book is the fruit of over a decade of work by two renowned researchers in the field of voice studies, Jody Kreiman of the UCLA Department of Head and Neck Surgery and Diana Sidtis (formerly van Lancker) of the NYU Steinhardt School of Culture, Education and Human Development. The pairing of these two professors from disparate corners of the discipline (even more disparate when one considers that Sidtis’ reputation was first established in the field of neurology) yields a book of remarkable breadth, along with great depth. Unlike most authors of books about the voice, who stay firmly within the confines of one discipline or
another, Kreiman and Sidtis address a full spectrum of voice-related issues from neurology, vocal anatomy, linguistics, acoustics, psychology, vocal performance, forensics, and other disciplines. Their thoughtful discussion of vocal plasticity includes a detailed acoustic analysis of the special artistry of Mel Blanc, who was the voice of Bugs Bunny, Porky Pig, Elmer Fudd, Sylvester the Cat and Tweety Bird.

There are ten chapters in the book, plus an extensive bibliography and two indices. Chapter 1 (pp. 1-24) is an introduction to voice quality, including vocal taxonomies from Roman times to the present. Chapter 2 (pp. 25-71) is a detailed description of the vocal articulators, with a great deal more attention paid to the characteristics of the vocal source (vocal folds, laryngeal cartilages and musculature) than can be found in most linguistics texts, even in the oeuvre of the late Peter Ladefoged, to whom the authors dedicate their work. Chapter 3 (pp. 72-109) covers the neurological foundations of voice, including audition, another topic often excluded from linguistics texts. Chapter 4 (pp. 110-155) surveys the range of a speaker’s physical characteristics, including age, size, gender, and race, which have been associated with listeners’ judgments of the voice. (Perhaps surprisingly, the survey of singing voices, which are commonly associated with body type – tenors presumed to have short necks, for example – is relegated to the Miscellany in chapter 10.) Chapters 5 (pp. 156-188) and 7 (pp. 237-259) discuss forensic applications of voice studies. Chapter 6 (pp. 189-236) returns to neurolinguistics, this time examining the brain mechanisms that underlie voice processing. This allows the authors to discuss their own groundbreaking work on familiar and unfamiliar voices, and to promote the well-supported, but novel, theory that familiar voice recognition is of prime importance to humans (and other species) because it is evolutionarily advantageous, while unfamiliar voice recognition, so central to jurisprudence, has few practical applications, and thus relatively little importance, in biological and social functioning. Chapter 8 (pp. 260-301) considers the linguistic applications of vocal features, such as prosody. Chapter 9 (pp. 302-360) delves into paralinguistic qualities, such as emotion and personality. Chapter 10 (pp. 361-397) is hardly the afterthought that the title “Miscellany” implies, instead encompassing voice in law enforcement, media, and singing.

Zeki Majeed Hassan and Barry Heselwood (eds.) (2011)
Instrumental Studies in Arabic Phonetics

Reviewed by
Judith Rosenhouse
SWANTECH Ltd. Haifa Israel

The Arabic language has been increasingly gaining researchers’ interest in its structure and development. This reviewed book is concerned with the instrumental study of Arabic phonetics. It has been edited in collaboration by two researchers from different European universities and includes 15 chapters written
by researchers from diverse locations. The book includes an Introduction, and four parts. Following the papers, there is an appendix describing the technical details of the instruments used in the studies (pp. 356-358) and a subject index (pp. 359-365). It should be noted that all the papers include many tables, figures and graphs, partly with colors, which present the advantages of each instrument and make the results more vivid.

The Introduction, written by the editors, B. Heselwood and Z. M Hassan (pp. 1-25) is comprised of three sections: a brief introduction, historical perspective, and a discussion of the contribution in this volume. This chapter is a good introduction for readers who are not acquainted with the historical background of the Arabic language in general and phonetics in particular, the issues that were discussed in the Classical period of Arabic, including methodological issues, and a survey of the main trends in the modern phonetic studies of Arabic. The papers of the volume are described from the point of view of subjects and method of work, e.g., studies advocating a particular phonological perspective (such as quantity, consonantal assimilation, or guttural laryngeals) or focusing on realisational phenomena (such as airflow, measuring of plain and emphatic coronals by nasoendoscopic, videofluoroscopic and acoustic means, coarticulatory effects measured by electromagnetic articulographic devices etc.). The references to this Introduction are divided into Arabic and other classical text editions, and modern sources.

Part I contains three chapters:

Adamantios I Gafos, Philip Hoole and Chakir Zeroual authored “Preliminary study of Moroccan Arabic word initial consonant clusters and syllabification using electromagnetic articulography” (pp. 29-46). Consonantal clusters (which may include 3-4 consonants) characterize Moroccan Arabic speech, and have been a frequent topic of phonetic research. This paper examines this issue using electromagnetic articulography in word initial clusters, in particular in the context of syllable structure and stability. The authors define simplex and complex onset alignments which differ in duration. Word triads with 1-3 word initial consonants of two speakers of the Moroccan Oujda dialect was recorded using the articulograph. The results of interval durations as a function of cluster size showed significant differences for all comparisons (C vs. CC, CC vs. CCC; C vs. CCC), which support the simplex onset hypothesis in these data. The phonetic heuristic for simplex onsets is right-edge-to anchor interval stability, and across different sizes, this interval was more stable than left-edge-to anchor or centre-to-anchor intervals.

Zeki Majeed Hassan wrote “An acoustic phonetic study of quantity and quantity complementarity in Swedish and Iraqi Arabic (pp. 47-62). Phonologically, vowel and consonant length is distinctive in Arabic. Segmental duration in a CV context is, however, not robust. Swedish distinguishes vowel durations also, but its behavior in this context appears to differ from Arabic. This state triggered the present study. After description of Swedish and Arabic phonetic behavior in some detail, the aim of this study is defined as seeing if the durational differences operate differently in long vowels from those in which short vowels precede the consonant. The informants (a speaker of Swedish and some speakers of Arabic from Hassan 1981) were recorded and their vowel and consonant durations were measured and statistically analyzed. The study showed that durational differences of V:C vs. V:C: were significantly different from those of VC vs. VC: in Iraqi Arabic (and different from Swedish).
Barry Heselwood, Sara Howard and Rawya Ranjous presented “Assimilation of /l/ to /r/ in Syrian Arabic: An electropalatographic and acoustic study” (pp. 63-98). The subject refers to the assimilation of word final /l/ to word initial /r/ in Syrian Arabic, which involves manner of articulation differences in such contexts. The authors note that such processes were described already in Sibawayhi’s “Book” (1967) from about 1200 years ago as *idγa:m* (they discuss it here in relation to increasing sonority). They also mention that manner changes are apparently less researched than place articulation changes. The speech of three women from Damascus, Homs and Latakaya (all holding PhD degrees, recorded in the UK) was examined using the electrographic instrument.

The findings indicate that in slow, normal and fast speech rates this assimilation is optional, and most common in the fast speech rate. [r] is shown to be member of both sets, but not [l]. The assimilation process is thus shown to be gradient and/or optional, which means that no assimilation of place or manner of articulation is categorical if there is a difference in the potential for the word-final consonant to be released.

In Part 2 there are also three papers:

Barry Heselwood and Feda Al-Tamimi study “The laryngeal and pharyngeal consonants in Jordanian Arabic using nasoendoscopy, videofluoroscopy and spectrography” (pp. 101-127). This paper describes the production of laryngeal and pharyngeal consonants by seven Jordanian speakers. As expected, the retraction of the epiglottis is shown by nasoendoscopy and videofluoroscopy to be important in distinguishing the two categories and spectrography shows the F1-F2 differences between them. Psychoacoustic analysis further shows F1 and F2 integration in open vowels in the context of pharyngeal, but not laryngeal consonants. Analyzing and discussing the findings lead the authors to suggest that Esling’s model (1999) of the laryngeal articulator can be regarded so that the pharyngeals /h, h/ and /ʕ, ʕ/ are proportional to /s-s/ and /d-d/ and the emphatic counterpart of the (plain) laryngeals or emphatic laryngeals.

Next is Kimary Shahin’s paper “A phonetic study of guttural laryngeals in Palestinian Arabic using laryngoscopic and acoustic analysis” (pp. 129-140). In this paper laryngeals [ʕ, h] are viewed together with velar [x, k] in a natural category of “gutturals.” The study is based on the speech of a male speaker of Palestinian Arabic recorded by a rhino-laryngeal-stroboscope connected directly to a digital Camcorder so that simultaneous audio and video recordings could be used. A similar patterning was found for Hebrew by Laufer and Baer (1988) and was presented again in this paper. Shahin suggests that this lack of pharyngeal (aryepiglottic) articulation in the two Semitic languages is a phonetics-phonology mismatch. (The author also comments concerning Heselwood & Al-Tamimi’s paper that the aryepiglottic articulation in laryngeals found in Jordanian Arabic is below the threshold of retraction that counts as pharyngeal articulation.) Shahin further considers this finding according to the Grounding Hypothesis (Archangeli & Pulleyblank, 1994) and Optimality Theory (Prince and Smolensky, 1993/2004) and concludes that the Arabic laryngeals have lost their Retracted Tongue Root specification and are implemented in the phonetics without aryepiglottic articulation, in violation of the Grounding Hypothesis, which in Arabic shows an epiphenomenon of constraint interaction.

Mohamd Yeou and Shinji Maeda’s paper “Airflow and acoustic modeling of pharyngeal and uvular consonants in Moroccan Arabic” (pp. 141-162) is next.

---

2 The speaker’s recorded dialect is the Palestinian Arabic dialect spoken in Jaffa/Yafo, Israel.
This paper examines the relation between calculated idealized models of the vocal tract area functions for \( /\ell, \hat{\ell}, \hat{\h}, \hat{\v}, \hat{\i}, \hat{\j}/ \) and real data of these consonants. Synthesized versions of \([aCa]\) consonant sequences are studied perceptually (with ten native Moroccan Arabic speakers). This study suggests that these consonants should be considered approximants rather than fricatives because values of supraglottal constriction and glottal constriction are estimated higher than in simple fricatives; in addition, the spectrograms show a vowel-like formant structure. Moreover, studying these vowels by calculated and measured air flow values yielded a higher range than for normal fricatives. This study (also) concludes with considering these consonants as a guttural group in line with McCarthy (1994) and that the stricture of these gutturals is important in differentiating them from pharyngealized (emphatic) consonants \((/s^\hat{\v}, d^\hat{\i}, t^\hat{\j}, q)/\) and the uvular stop \(/q/\).

The third part has more papers than the other parts. We find here:

Feda Al-Tamimi and Barry Heselwood’s paper “Nasoendoscopic, videofluoroscopic and acoustic study of plain and emphatic coronals in Jordanian Arabic” (pp. 165-191). Following the introductory summary of studies of Arabic emphatics. The authors present the study using the three instruments mentioned in the title. The initial, medial and final positions of the test words were studied for emphatics and corresponding plain consonants and their effects on adjacent vowels in urban Jordanian Arabic. The findings for the studied consonants \((s, s^\hat{\v}, t, t^\hat{\j}, d, d^\hat{\i})\) with the vowels (short and long /i, u, a/) are given in tables, pictures and spectrograms of the nasoendoscopic, videofluoroscopic and acoustic techniques. The analysis of the findings explains the complex patterns of articulatory adjustments involved in the process, which include many anatomical structure components (tongue, hyoid bone, pharyngeal wall, epiglottis, pyriform sinuses, laryngeal aditus, and larynx). They conclude by underscoring the importance of studying both the “plain” and emphatic coronals, because they are not distinct only by the features of emphasis.

Mohamed Embarki, Slim Ouni, Mohamed Yeou, Christian Guilleminot & Sallal Al-Maqtari wrote “Acoustic and electromagnetic articulographic study of pharyngealisation: Coarticulatory effects as an index of stylistic and regional variation in Arabic” (pp. 193-215). This study refers to emphatics, while considering four dialect varieties (DA) from Yemen, Kuwait, Jordan and Morocco, as well as Modern Standard Arabic (MSA). The authors consider the differences between phonemes and varieties from the point of view of Locus Equations (LEs, Lindblom, 1963) and further developed by others (e.g., Krull, 1989, Sussman et al. 1998). The material was based on recorded speech (24 MSA and 24 DA words) of 16 male adult speakers from each of those countries. The analyzed vowels were measured at their F2 onset and F2 mid positions, averaged for each dialect, and then analyzed for their LEs. The results are presented for the vowels in the pharyngeal and plain contexts for MSA and DA and the LE slopes. Inter-dialect differences were also calculated and presented in the paper. In a second experiment, data were acquired by the 3D articulograph for words spoken by a Tunisian male speaker. The words covered the plain and pharyngealised consonants in initial, middle and final positions. In conclusion, the authors stress the relevance of LE parameters for distinction of pharyngealized and plain consonants and stylistic (MSA vs. DA) and regional indexation. The slope values for pharyngealized consonants are flatter in MSA compared to DA, and steeper for the plain consonants in MSA compared to DA. The distance between pharyngealized and plain consonants is shallower in DA compared to MSA.
These and additional differences between MSA and DA can be based on cognitive differences due to the more careful speech of MSA compared with DA. Differences are also evident between the DA data, so that the LE slopes “divide” them into Eastern vs. Western groups, which is in line with other known features of this classification. The LE slopes of DA suggest that the dialects are developing in a direction of weakening of pharyngealization, and probably to merging with their plain cognates. In any case, many coarticulatory patterns are apparently involved in Arabic pharyngealized consonants, and fine gestural adjustments and effects varying according to style and regional variety.

Zeki Majeed Hassan and John H. Esling authored the paper “Investigating the emphatic feature in Iraqi Arabic: Acoustic and articulatory evidence of coarticulation” (pp. 217-234). The study begins with a survey of the literature on the effects of emphatic consonants and directionality. The study then examines differences between emphatic and pharyngeal consonants as pronounced by two speakers of Iraqi Arabic and analyzed them using acoustic analysis and direct visual articulation by transnasal laryngoscopy. The findings indicate that in Iraqi Arabic, emphatic articulation differs from pharyngeals in the shape of the pharynx. Laryngeal constriction with tongue retraction and larynx raising characterizes pharyngeals, but larynx height seems to be low in the secondary articulation of emphatics, with a narrow pharynx configuration and raised tongue dorsum. They also find that emphatic consonant effects spread forward and backward, but inconsistent data may be affected by pitch contours and stress, and the emphatic effects are blocked or modified under certain phonotactic conditions, such as segments like /ći, jı, Ĵı/. The authors also note that in contrast with their data, Al-Tamimi & Heselwood (this volume) found a raised larynx, which calls for further research. Hassan & Esling conclude by saying that these findings support the concept that the emphatic feature prevails over multisyllabic words in Arabic, that anticipatory coarticulation is language-specific and that the secondary articulation is assumed prior to the primary one and dominates the word.

Janet C.E. Watson and Alex Bellem provided “Glottalisation and neutralization in Yemeni Arabic and Mehri: An acoustic study” (pp. 23-256). This paper presents a little studied aspect of a Yemeni Arabic dialect, that of the capital San’a, and some interesting phonetic features of that dialect. The authors compare these features to Mehriyot another little studied dialect, which belongs to a different, but related language. The authors thus bring the reader into the realm of the Afroasiatic phylum, proto-Semitic Arabic, and contemporary Arabic dialects in Yemen. This country has numerous variegated Arabic dialects which differ from other dialects of the Arabian Peninsula (and others) in being nearer to South Arabian languages, of which Mehriyot is one which still exists in Yemen. The studied features, pre- and post-glottalisation in word (or utterance) final position, are demonstrated by recorded examples of the speech of a teenage girl and an adult male speaker. The examples show clearly ejective final glottalization following /g/ (which becomes /k’/), /d/ (which becomes /h’/) and /t/ which becomes /t’/ and the aspirated release of pre-pausal /t/ > [tʰ]. In addition, the authors describe the release of pre-pausal fricative in /ʃ/ > /ʃ/ and /ʃ/, /ʃ’/ > /ʃ/ , /b/ > /b’, /d/ > /d’, /t/ > /t’/ even total deletion of /b’, d’, t’/. Similar processes occur in Mehriyot, but not under the same phonetic conditions. In general, all San’a Arabic sonorants are subject to glottalisation prepaussally, while /m, n, r, l/ are devoiced and preglottalised and the coronals are often lenited to the point of elision, especially noticeable with /n/. Vowels are post-glottalised, but never elided. The description is based on a few speakers’ speech due to
technical reasons. The features were frequent in other speakers, and thus the authors call for more research on this subject. Another interesting aspect of this feature which the authors intend to research is the sociolinguistic difference noted between male and female speakers’ manner of glottalisation.

The next paper is by Bushra Adnan Zawaydeh and Kenneth de Jong, “The phonetics of localizing uvularisation in Ammani-Jordanian Arabic: An acoustic study” (pp. 257-276). This paper examines contrasts involving consonant uvularisation as found in the speech of Jordanian Arabic, with a focus on how the degree of uvularisation across a word containing a uvularised consonant occurs. These authors distinguish uvularisation from pharyngealisation, since the former is expressed mainly in lowered F2, while the latter is expressed in raised F1. Three recorded word lists (corpora) spoken by three male and three female speakers from Amman, the capital of Jordan, were examined acoustically and statistically. The variables were: the trigger (no trigger, an emphatic, or /q/) in the first two lists, and the distance of the target from the trigger measured in the number of vowels between them, the direction of the target, and the presence of the trigger in the third list. Corpus I results revealed robust effects of both trigger and blocker, and all three categories of triggers (plain, uvulars, emphatics) differed significantly from one another. Corpus 2 results indicated an effect of trigger and a weak effect of blocker. Together, these two corpora established a sizeable phonetic effect of emphatics extending in both directions (left and right) regardless of a blocking segment. For corpus 3, the distance had a significant effect on F2 but direction did not, although there was an interaction between direction and distance. The average F2 values seemed to be similar to those found for perseveratory uvularisation from /q/ in corpus 1. Four general effects in uvularisation associated with emphatic consonants were found: 1. The most uvularised vowels (lowest F2) immediately followed the emphatic trigger. 2. As the vowel is more distant following the consonant, uvularisation amount weakens considerably. 3. Anticipatory spreading is not sensitive to distance. 4. the degree of uvularisation is very strong 2-3 vowels before the trigger but becomes less towards the trigger, thus F2 is higher in the 2nd vowel before the emphatic consonant than in vowels earlier in the word. These findings are discussed with the conclusion that more research is needed to understand what makes such readily detectable typological variation in uvular articulation.

Chakir Zeroul, John H. Esling and Philip Hoole presented “EMA, Endoscopic, ultrasound and acoustic study of two secondary articulations in Moroccan Arabic: labial velarisation vs. emphasis” (pp. 277-297). This paper describes the use of electromagnetic, articulography, endoscopy and ultrasound in research and finds the following: the studied Moroccan Arabic emphatic stops can be considered pharyngealized; the emphatics /tʰ/, /dʰ/ are not velarized, and /d̪ʰ/ has a slight labialization. They also find that Moroccan Arabic geminate labials are velarized (not emphatic or pharyngealized). They confirm that labialisation in word initial clusters /#C1Cw/ (where C1 or C2 is a velar or uvular consonant) is associated with C1, even if the latter is a coronal. Velar and uvular labialized dorsals, e.g., /kʷ/, /gʷ/, /qʷ/) are also more retracted than their non-labialised counterparts. These findings are based on recording one speaker by electromagnetic articulography and nasendoscopy, and by recording two speakers’ by ultrasound. The acoustic measures of F1, F2 and F3 were measured from five tokens at the onset, midpoint and offset of the first vowel /a/ of the EMA items. Several paired t-tests were performed separately for each pair of the segment. The findings are partly in line with the close relation reported between
labial and dorsal gestures (e.g. Ladefoged & Maddieson, 1996) since Moroccan Arabic labialized labials are produced with labial velarisation and labialized dorsals are produced with a more retracted tongue dorsum position combined with labialization.

The last, fourth part includes a paper on intonation and a paper on acquisition:

Sam Hellmuth wrote “Acoustic cues to focus and givenness in Egyptian Arabic” (pp. 301-324). The dialect discussed here is that of Cairo, the largest city in Egypt. The paper deals with focus, which indicates a choice between alternatives, similar to contrastive focus, and givenness denotes availability or prior mention in the discourse. Focus and givenness are expressed in various manners in different languages, e.g., intonational pitch accents or deaccentuation. (e.g., “was that a blue car? No, it was a RED car” where RED is in focus and the 2nd time ‘car’ is given) In Egyptian Arabic, the latter has been observed not to occur (Hellmuth, 2005). However, the literature of Arabic intonation mentioned graded variation in F0 excursion in some Arabic dialects. The author tests whether intensity can replace pitch in marking focus or givenness in this dialect by measuring general intensity, its spectral tilt, as well as F0 and duration, using three male and three female speakers’ recorded data set of sentences with the structure of VOO (Verb - Object direct - Object indirect). The results show a similar basic pattern to the SVO dataset (Hellmuth, 2006a): although there is great speaker variation, there is a trend towards greater FO excursion on focused items and compressed F0 excursion after such an item and this trend is not matched in changes in duration. Also, no intensity variation was observed with the givenness status of an object. This finding is discussed in the context of English intonation.

Ghada Khattab offered “Acquisition of Lebanese Arabic and Yorkshire English /l/ by bilingual and monolingual children: a comparative spectrographic study” (pp. 325-354). This last chapter presents and discusses auditory and acoustic analysis of the production of /l/ in 23 English and Arabic monolingual and bilingual children and adults. The children were in three age groups (5, 7, 10 years old). In both languages, /l/ has different varieties (‘clear’ and ‘dark’ as referred to in English and ‘emphatic’ for Arabic), but their use conditions differ. The findings suggest that these two languages differ not only in their phonotactic rules as to the occurrence of these variants, but also as to the actual realization of clear /l/. The two languages seem to be similar, however, in their having pharyngealisation in their ‘dark’ variant. The results are averaged and analyzed statistically and presented by their F2 and F1 values for both languages in Hz and in Bark values. The bilinguals’ parents seem to have retained the clear Arabic [l] in the coda position even after years of living in UK. The low occurrence of /l/ in Arabic did not allow clear conclusions, but F2 was lower and F1 was higher, in parallel with English dark /l/. But the bilingual children showed no significant difference between F2 frequencies for the English and Arabic clear [l], which were closer to those of monolingual English patterns than to Arabic ones. The author attributes this result to the dominance of English in their speech and notes their control of clear /l/ in each language.

In sum, this is a collection of papers that make an important, interesting and high quality contribution to the field of phonetics. The papers study major issues in this field (e.g., pharyngealisation, laryngealisation, emphasis, intonation). The studies also represent a nice collection of Eastern and Western Arabic dialects (e.g., Moroccan, Egyptian, Yemenite, Iraqi, Syrian, Jordanian, and Israeli), as well as some non-Arabic languages (Swedish, Mehri, English) and up-to-date
methods of phonetic research with modern research instruments. Most of the Arabic words are written in phonetic transcription (not only IPA). The lists of references at the end of each paper are rich and instructive. All the authors of these papers have previously contributed to Arabic acoustic phonetics, and some of them collaborate in these chapters. It may be hoped that such studies will continue to enrich our knowledge. All these factors make this volume an invaluable addition to the phonetic book-shelf for scholars of both Arabic and general phonetics.

References
MEETINGS, CONFERENCES AND WORKSHOPS

2011

12-14 January 2011
CUNY Conference on the Phonology of Endangered Languages
Manhattan, NY, USA
http://cunyphonologyforum.net/endan.php

20-22 January 2011
8th Old World Conference in Phonology (OCP8)
Marrakech, Morocco

28-31 January 2011
New Tools and Methods for Very-Large-Scale Phonetics Research (VLSP 2011)
Philadelphia, PA, USA.
http://www.ling.upenn.edu/phonetics/workshop/

11-12 April 2011
BACL 2011, 3rd Colloquium of the British Association of Clinical Linguistics
Leeds, UK
http://www.leedsmet.ac.uk/hss/subject_groups_0538C63C6F7C4552A9B6352602B573AE.htm

16-18 May 2011
Speaking 2011: Speaking in a Foreign language - Effective Learning, Teaching and Assessment
Konin, Poland
http://sites.google.com/site/konferencjamowienie2011/english

27-28 May 2011
TRANSCRIBING, WRITING, FORMALISING - 2, 25th international conference
Université d'Orléans, France
http://www.mshs.univ-poitiers.fr/cerlico/cerlico.htm

9-11 June 2011
6th International Conference on Speech Motor Control
Groningen - Nijmegen, NL
http://www.slp-nijmegen.nl/smc2011
13-15 June 2011

**Phonetics without Borders (PhwB-2011)**
Blagoveshchensk, Amur Region, Russia

16-17 June 2011

**International Child Phonology Conference**
University of York, UK

17-21 June 2011

**Workshop on Phonetic Grounding in English Phonology**
Boston, MA, USA

21–22 June 2011

**Phonetics and Phonology in Iberia 2011 (PaPI 2011)**
Tarragona, Spain
http://www.urv.cat/deaa/PaPI2011/home.html

27 June-1 July 2011

**Forum Acusticum 2011: European conference of the European Acoustics Association**
Aalborg, Denmark
http://www.fa2011.org/

30 Jun - 02 Jul 2011

**French Phonology Network Annual Meeting, RFP 2011**
Tours, France

16 August 2011

**Coarticulation in New Varieties of English. A satellite event to ICPhS XVII.**
Hong Kong SAR, China
http://www.reading.ac.uk/epu/ICPhS17_Satellite/

17-21 August 2011

**The 17th International Congress of Phonetic Sciences (ICPhS XVII)**
Hong Kong SAR, China.
http://www.icphs2011.hk

24-26 August 2011

**Speech and Language Technology in Education (SLaTE) 2011**
Venice, Italy
http://project.cgm.unive.it/events/SLaTE2011/
27-31 August 2011
   **Interspeech 2011**
   Florence, Italy
   http://www.interspeech2011.org/

12-14 September 2011
   **The Prosody-Discourse Interface**
   Salford, Manchester, UK
   http://www.famss.salford.ac.uk/page/pdi_conference

7-8 October 2011
   **Pronunciation in Second Language Learning and Teaching (PSLLT): The Confluence of Social Factors and Pronunciation:Accent, Identity, Irritation and Discrimination**
   Ames, IA, USA
   http://pslltconference.com

2012

1-2 March 2012
   **The Phonology of Contemporary English: Variation and Change (PAC 2012)**
   Toulouse, France
   http://w3.pac.univ-tlse2.fr

23-24 March 2012
   **3rd Belgrade International Meeting of English Phoneticians (BIMEP 2012)**
   Belgrade, Serbia

26-28 March 2012
   **Colloquium of the British Association of Academic Phoneticians**
   University of Leeds
   http://www.personal.leeds.ac.uk/~lnpbaap/

27 March 2012
   **GLOW workshop on Prosodically-Coded Information Structure**
   Potsdam, Germany
   http://www.ling.uni-potsdam.de/~glow

29 March 2012
   **L2 Prosody Workshop**
   Bangor, Wales, United Kingdom
   http://www.bilingualism.bangor.ac.uk/events/L2_Prosody_Workshop.php.en

17–20 April, 2012
   **The 10th International Conference on the Computational Processing of Portuguese**
   http://www.propor2012.org/
22-24 April 2012
   2nd International Phonetics & Phonology Conference Shanghai
   Shanghai, China

2-3 May 2012
   The Listening Talker: An interdisciplinary workshop on natural and
   synthetic modification of speech in response to listening conditions
   University of Edinburgh, UK
   http://listening-talker.org/workshop

2-4 May 2012
   2nd Workshop on Sound Change
   Kloster Seeon, Bavaria, Germany
   http://www.phonetik.uni-muenchen.de/institut/veranstaltungen/soundchange/

4-5 May 2012
   7th North American Phonology Conference (NAPhC7)
   Montreal, Quebec, Canada
   http://linguistics.concordia.ca/naphc7/

22-25 May 2012
   6th International Conference, Speech Prosody 2012. Prosody in the real
   world: Understanding and approaching human prosodic performance
   Shanghai, China
   http://www.speechprosody2012.org/

27-29 May 2012
   International Symposium on Tonal Aspects of Languages (TAL)
   Nanjing, Jiangsu Province, China
   http://www.TAL2012.org

29 May 2012
   Discourse Coherence and Prosody (CDP2012)
   Lille, France
   http://evenements.univ-lille3.fr/je-cdp

4-6 June 2012
   International Child Phonology Conference (ICPC) 2012
   Twin Cities, Minneapolis, MN, USA
   http://www.tc.umn.edu/~munso005/ChildPhonology

25-28 June 2012
   Odyssey 2012: The Speaker and Language Recognition Workshop. An
   ISCA Tutorial and Research Workshop
   Singapore
   http://www.odyssey2012.org/
27-30 June 2012
14th International Clinical Phonetics & Linguistics Association (ICPLA) Conference
Cork, Ireland
http://www.icpla2012.com/

2 July 2012
Teaching and Learning Pronunciation: Local and global perspectives on research and practice
Cairns, Australia

19-21 July 2012
Perspectives on Rhythm and Timing (PoRT)
Glasgow, United Kingdom
http://www.gla.ac.uk/rhythmsscotland/

27-29 July 2012
13th Conference on Laboratory Phonology (LabPhon13)
Stuttgart, Germany
http://www.labphon13.labphon.org/

15-17 August 2012
Nordic Prosody XI.
Tartu, Estonia
http://www.nordicprosody.ut.ee/

3-5 September 2012
International Symposium on Imitation and Convergence in Speech - ISICS 2012
Aix-en-Provence, France
http://spim.risc.cnrs.fr/ISICS.htm

5-7 September 2012
Advances in Visual Methods for Linguistics – AVML
York, United Kingdom
http://avml2012.wordpress.com/

9-13 September 2012
InterSpeech 2012
Portland, Oregon, USA
http://interspeech2012.org

12-14 October 2012
22nd Japanese/Korean Linguistics Conference
Tokyo, Japan
http://www.ninjal.ac.jp/jk2012/
CALL FOR PAPERS

The *Phonetician* will publish peer-reviewed papers and short articles in all areas of speech science including articulatory and acoustic phonetics, speech production and perception, speech synthesis, speech technology, applied phonetics, psycholinguistics, sociophonetics, history of phonetics, etc. Contributions should primarily focus on experimental work but theoretical and methodological papers will also be considered. Papers should be original works that have not been published and are not being considered for publication elsewhere.

Authors should follow the *Journal of Phonetics* guidelines for the preparation of their manuscripts. Manuscripts will be reviewed anonymously by two experts in phonetics. The title page should include the authors’ names and affiliations, address, e-mail, telephone, and fax numbers. Manuscripts should include an abstract of no more than 150 words and up to four keywords. The final version of the manuscript should be sent both in .doc and in .pdf files. It is the authors’ responsibility to obtain written permission to reproduce copyright material.

All kinds of manuscripts should be sent in electronic form (.doc and .pdf) to the Editor. We encourage our colleagues to send manuscripts for our newly released section entitled MA research, which is a summary of the students phonetics research describing their motivation, topic, goal, and results (no more than 1,200 words).

INSTRUCTIONS FOR BOOK REVIEWERS

Reviews in the *Phonetician* are dedicated to books related to phonetics and phonology. Usually the editor contacts prospective reviewers. Readers who wish to review a book mentioned in the list of “Publications Received” or any other book, should address the editor about it.

A review should begin with the author’s surname and name, publication date, the book title and subtitle, publication place, publishers, ISBN numbers, price, page numbers, and other relevant information such as number of indexes, tables, or figures. The reviewer’s name, surname, and address should follow “Reviewed by” in a new line.

The review should be factual and descriptive rather than interpretive, unless reviewers can relate a theory or other information to the book which could benefit our readers. Review length usually ranges between 700 and 2500 words. All reviews should be sent in electronic form to Prof. Judith Rosenhouse (e-mail: swantech@013.net).
ISPHS MEMBERSHIP APPLICATION FORM

Please mail the completed form to:

Treasurer:
Prof. Dr. Ruth Huntley Bahr, Ph.D.
Treasurer’s Office:
Dept. of Communication Sciences and Disorders
4202 E. Fowler Ave. PCD 1017
University of South Florida
Tampa, FL 33620 USA

I wish to become a member of the International Society of Phonetic Sciences

Title: _____ Last Name: ____________________ First Name: ___________________
Company/Institution: _______________________________________________________
Full mailing address: ________________________________________________________
Phone: ____________________________ Fax: ____________________________
E-mail: ___________________________________________
Education degrees: __________________________________________________________
Area(s) of interest: __________________________________________________________

The Membership Fee Schedule (check one):
1. Members (Officers, Fellows, Regular) $ 30.00 per year
2. Student Members $ 10.000 per year
3. Emeritus Members NO CHARGE
4. Affiliate (Corporate) Members $ 60.000 per year
5. Libraries (plus overseas airmail postage) $ 32.000 per year
6. Sustaining Members $ 75.000 per year
7. Sponsors $ 150.000 per year
8. Patrons $ 300.000 per year
9. Institutional/Instructional Members $ 750.000 per year

Go online at www.isphs.org and pay your dues via PayPal using your credit card.

☐ I have enclosed a cheque (in US $ only), made payable to ISPhS.

Date _____________________ Full Signature _______________________

Students should provide a copy of their student card
NEWS ON DUES

Your dues should be paid as soon as it convenient for you to do so. Please send them directly to the Treasurer:

Prof. Ruth Huntley Bahr, Ph.D.
Dept. of Communication Sciences & Disorders
4202 E. Fowler Ave., PCD 1017
University of South Florida
Tampa, FL 33620-8200 USA
Tel.: +1.813.974.3182, Fax: +1.813.974.0822
e-mail: rbahr@usf.edu

VISA and MASTERCARD: You now have the option to pay your ISPhS membership dues by credit card using PayPal if you hold a VISA or MASTERCARD. Please visit our website, www.isphs.org, and click on the Membership tab and look under Dues for the underlined phrase, “paid online via PayPal.” Click on this phrase and you will be directed to PayPal.

The Fee Schedule:
1. Members (Officers, Fellows, Regular) $ 30.00 per year
2. Student Members $ 10.00 per year
3. Emeritus Members NO CHARGE
4. Affiliate (Corporate) Members $ 60.00 per year
5. Libraries (plus overseas airmail postage) $ 32.00 per year
6. Sustaining Members $ 75.00 per year
7. Sponsors $ 150.00 per year
8. Patrons $ 300.00 per year
9. Institutional/Instructional Members $ 750.00 per year

Special members (categories 6–9) will receive certificates; Patrons and Institutional members will receive plaques, and Affiliate members will be permitted to appoint/elect members to the Council of Representatives (two each national groups; one each for other organizations).

Libraries: Please encourage your library to subscribe to The Phonetician. Library subscriptions are quite modest – and they aid us in funding our mailings to phoneticians in Third World Countries.

Life members: Based on the request of several members, the Board of Directors has approved the following rates for Life Membership in ISPhS:

- Age 60 or older: $ 150.00
- Age 50–60: $ 250.00
- Younger than 50 years: $ 450.00