Workshop
Production and Comprehension of Conversational Speech

12 - 13 December 2011

Organizers:
Mirjam Ernestus
Iris Hanique
Natasha Warner

Sponsored by:
European Young Investigator Award to Mirjam Ernestus
Max Planck Institute for Psycholinguistics

MONDAY, December 12		
9:25-9:30	Welcome	
9:30-11:00	***Session 1***	
	Laurence White, Lukas Wiget, Katharine Barden, Ahsanul Kabir, Olesya Rauch & Sven L. Mattys	"Turn left at the grey tanker": Production and perception of segmentation cues in spontaneous speech
	Vincent Aubanel, Julian Villegas, Martin Cooke	Conversing in the presence of another conversation: interactive and Lombard effects
	Katharine Barden, Sarah Hawkins	Towards ecological validity in studying adaptation to accents: an investigation of the role of morphological structure in perceptual learning
11:00-11:30	Coffee break	
11:30-12:30	***Session 2***	
	Marco van de Ven, Benjamin V. Tucker, Mirjam Ernestus	Semantic context effects in the processing of unreduced and reduced sentences
	Richard Ogden, Beatrice Szczepek-Reed	Phonetic details as interactional resources

12:30-13:30	Lunch	
13:30-14:30	***Session 3***	
	Benjamin V. Tucker, Antti Arppe	Allophonic realizations of the phoneme /t/ in an English spontaneous speech corpus
	Tyler Kendall, Ann R. Bradlow, Brett Margolis	Speech timing and accommodation in native and non-native English interactions
14:30-15:00	Coffee break	
15:00-16:30	***Session 4***	
	Oliver Niebuhr, Evelin Graupe, Laura C. Dilley	You don't have to say a word – How duration and F0 trigger or hinder the perception of function words in German
	Philip Dilts, R. Harald Baayen, Benjamin V. Tucker	Describing and predicting phonetic reduction in a corpus of spontaneous speech
	Valerie Hazan, Rachel Baker	Between-speaker variability in the clarity of spontaneous speech
16:30	Excursion and dinner	

TUESDAY, December 13		
8.30-9.00	Coffee	
9:00-10:30	***Session 5***	
	James M Scobbie, Eleanor Lawson, Jane Stuart-Smith	Tongues in conversation: a sociophonetic comparison of some Scottish lingual variables in conversational and wordlist speech
	Paul Carter, Leendert Plug	The phonetics of self-repair in spontaneous Dutch
	R. Harald Baayen, Benjamin V. Tucker	Acoustic duration is predicted by syntactic prototypicality
10:30-11:00	Coffee break	
11:00-12:00	***Session 6***	
	L. Ann Burchfield, Ann R. Bradlow	Comparison of reduction in spontaneous Mandarin and English
	Mybeth Lahey, Mirjam Ernestus	Acoustic reduction in infant-directed speech
12:00-13:00	Lunch	

13:00-14:00	***Session 7***	
	Susanne Brouwer, Holger Mitterer, Falk Huettig	Discourse context and the recognition of reduced and canonical forms
	Katja Pöllmann, Hans Rutger Bosker, James M. McQueen, Holger Mitterer	Do listeners form expectations about a speaker's tendencies to reduce?
14:00-15:30	Coffee and poster presentations	
15:30-16:30	***Session 8***	
	Cedric Gendrot, Martine Adda-Decker, Carolin Schmid	F0 declination in French: Broadcast news versus spontaneous speech
	Ann R. Bradlow, Valerie Hazan, Midam Kim, Michele Pettinato	Interlocutor accommodation across communication barriers in naturalistic conversations
16:30	Drinks	

"Turn left at the grey tanker": Production and perception of segmentation cues in spontaneous speech

Laurence White[1], Lukas Wiget[2], Katharine Barden[3],
Ahsanul Kabir[4], Olesya Rauch[3], Sven L. Mattys[3]
[1]*School of Psychology, University of Plymouth, USA*
[2]*Seminar für Allgemeine Sprachwissenschaft, Universität Zürich, Switzerland*
[3]*School of Experimental Psychology, University of Bristol, UK*
[4]*School of Computing and Mathematical Sciences, University of Greenwich, UK*

laurence.white@plymouth.ac.uk, lukas.wiget@uzh.ch, k.barden@bristol.ac.uk, A.Kabir@gre.ac.uk, psxop@bris.ac.uk, sven.mattys@bristol.ac.uk

Segmentation research asks how listeners locate word boundaries in the speech stream. Multiple cues – lexical, segmental, prosodic – have been shown to affect segmentation, but previous studies have almost exclusively used speech stimuli elicited in careful readings rather than natural communicative contexts. We report development of a segmentation-oriented corpus of spontaneous speech, the *Bristol Speech Corpus*, and assess how the production and perception of cues to word boundaries are modulated by the communicative situation.

Background. Conversational speech tends to be highly contextualised, with the production and interpretation of utterances dependent on a quasi-mutual understanding of the foregoing interaction. For example, speakers' degree of articulatory effort – hyperarticulation vs hypoarticulation – has been held to vary as a function of current communicative demands. This suggests that phonetic segmentation cues may be reduced where conversational context guides listeners' interpretation.

The Bristol segmentation map task and the Bristol Speech Corpus. We developed a map task to elicit spontaneous speech while controlling the occurrence of word-boundary cues. Pairs of speakers interacted conversationally regarding routes around stylised landmarks. Landmark names were one- or two-word phrases which modulated potential segmentation cues in five experimental conditions.

- Presence vs absence of word boundary: e.g., *paperback* vs *kipper bag*.
- Cross-boundary homophony: e.g., *great anchor* vs *grey tanker*.
- Word-initial vs non-initial stress: e.g., *cream* ***ˈrick****shaw* vs *cream re****ˈcluse***.
- Within-word frequency of cross-boundary diphone: e.g., *crea****m r****ickshaw* (low frequency) vs *dra****b r****ickshaw* (high frequency).
- Semantic predictability: e.g., *oil tanker* vs *seal tanker*.

To avoid speakers reading landmark names, they were pre-trained to recognise the landmark symbols. Utterances from the spontaneous map corpus were transcribed and subsequently re-recorded as read sentences by the same speakers, allowing comparison between two speech styles.

Production and perception of segmentation cues in the Bristol Speech Corpus. Analysis of the near-homophonous phrases and non-ambiguous controls showed the expected lengthening of consonants in word-initial compared with word-final position (e.g., [t] in *grey tanker* vs *great anchor*). However, the word-initial lengthening effect was significantly attenuated in map task tokens compared with those from read speech. Furthermore, perceptual data indicated that near-homophonous tokens from spontaneous speech became more ambiguous with repetition. Taken together, these results point to hypoarticulation of

segmentation cues in spontaneous speech, and suggest that previous research may overstate the importance of segmental and prosodic cues to word boundaries. Further perceptual experiments using the corpus – for example, examining the interpretation of phonotactic cues – support this conclusion.

Conversing in the presence of another conversation: interactive and Lombard effects

Vincent Aubanel, Julian Villegas, Martin Cooke

Ikerbasque & Language and Speech Laboratory, University of the Basque Country, Vitoria, Spain

v.aubanel@laslab.org, j.villegas@laslab.org, m.cooke@ikerbasque.org

Conversational speech is usually characterized by well described departures in production from carefully pronounced speech, while it maintains a fair level of comprehension. Less is known however on the speech production modifications induced by a background conversation on a foreground conversation, and how speakers manage to maintain intelligibility and comprehension in this scenario.

Extending a previous study with Spanish native speakers, we recorded pairs of English native speakers engaging in natural dialogues in the absence or presence of another talker pair. We observed small but significant increases in energy and F1 across conversations, and larger prosodic effects (increase of f0, decrease in speech rate) within dialogues, which are not easily explained in purely energetic masking terms. Indeed, background conversations are different from noise used in traditional Lombard studies in that they consist of intelligible speech, which creates the potential for informational masking at the ears of the interlocutor.

We further tested whether speakers' eye-contact with their interlocutor could influence their capacity to cope for the disruptive effect of a background conversation. Preliminary result point to an attenuation of Lombard effects (i.e., intensity, f0, F1, speech rate) in the absence of eye-contact condition, which could be the sign of an active monitoring of the informational content of the conversation, thus further minimizing the masking effect of the background conversation.

Towards ecological validity in studying adaptation to accents: an investigation of the role of morphological structure in perceptual learning

Katharine Barden[1], Sarah Hawkins[2]

[1]*School of Experimental Psychology, University of Bristol, UK*
[2]*Centre for Music and Science, University of Cambridge, UK*

k.barden@bristol.ac.uk, sh110@cam.ac.uk

Familiarity with a talker or accent can facilitate speech perception, but the extent to which perceptual learning occurs in 'everyday' conversation is not clear. Studies of perceptual learning tend to use unnatural stimuli (e.g. synthetic speech, isolated words) and unnatural tasks (e.g. lexical decision, phoneme categorisation) to familiarise listeners and assess their learning, and they focus primarily on whether listeners can learn to associate novel phonetic characteristics with low-level units such as features or phonemes. They thus neglect phonetic variation at other levels of representation in conversational speech (e.g. prosodic structure, grammatical function). As a step towards understanding accent adaptation in ordinary listening situations, the present experiment used relatively natural stimuli and tasks to test the hypothesis that listeners adapt to phonetic information that is systematically associated with the morphological structure of words.

The experiment comprised familiarisation and assessment phases. During familiarisation, subjects heard ten short stories read by a phonetician with a Standard Southern British English (SSBE) accent. Two versions of each story were recorded. 56 Control subjects heard the version in which all *re-* prefixes were realised with the standard SSBE [ri:]; and 56 Accent subjects heard the version in which all *re-* prefixes were realised as [rɪ] (e.g. [rɪθɪŋk] for *re-think*).

Perceptual learning was assessed using an intelligibility-in-noise task, comprising 18 experimental sentences and 60 fillers read by the same talker as the stories. Each experimental sentence included one keyword containing /ri:/, either as a Prefix or a NonPrefix (9 sentences each), and always realised as [rɪ]. There were no other /ri:/ syllables.
Prefix: *He aimed to **re-supply** the cocaine by Tuesday*
NonPrefix: *They claimed the **recent** violent campaign was stupid*
Subjects typed what they thought they heard. Keywords (e.g. *re-supply, recent*) were scored as correct or incorrect.

Results showed that Accent subjects learned to associate the [rɪ] pronunciation strongly with Prefix *re-*, and more weakly with NonPrefix /ri:/. Control subjects showed some learning during the assessment task, but their learning was specific to Prefix *re-*, despite equal exposure to the atypical pronunciation in Prefix and NonPrefix sentences. These results together suggest that the morphological status of an atypical pronunciation may affect its learnability, as well as confirming that very limited exposure can induce perceptual learning. We conclude that, while listening to stories, listeners monitor phonetic detail that reflects morphological structure, and can use this new knowledge to facilitate their understanding of sentences in difficult listening conditions.

Semantic context effects in the processing of unreduced and reduced sentences

Marco van de Ven[1], Benjamin V. Tucker[2], Mirjam Ernestus[3,1]
[1]*Max Planck Institute for Psycholinguistics, The Netherlands*
[2]*University of Alberta, Canada*
[3]*Radboud University Nijmegen, The Netherlands*

m.a.m.vande.ven@gmail.com, bvtucker@ualberta.ca, mirjam.ernestus@mpi.nl

In casual speech, words are often pronounced much shorter and with less articulatory effort than in careful speech. For example, the English words '*probably*' and '*yesterday*' may be realized like *proly* and *yesyay*. Listeners require contextual information to understand these highly reduced pronunciation variants (Ernestus, Baayen, and Schreuder, 2002). This raises the question which types of contextual information contribute to the recognition of reduced pronunciation variants, and to what extent. Previous research by Van de Ven, Tucker, and Ernestus (2010) has shown that, for isolated words, it takes longer until listeners can use semantic information from reduced variants to facilitate their processing of upcoming words. The present study investigates semantic priming effects in the processing of unreduced and reduced sentences, rather than isolated words, and thereby contributes to a better understanding of how humans deal with reduced speech in everyday listening situations.

To begin with, we conducted an auditory lexical decision experiment with semantic priming, investigating listeners' use of semantic contextual information in the comprehension of reduced and unreduced sentences. Each target sentence contained a prime and a target word (i.e. the sentence-final word), and these words were always semantically related to a certain extent. The semantic relatedness of each word pair was measured by means of Latent Semantic Analysis (Deerwester, Dumais, Furnas, Landauer, and Harshman, 1990). Participants were instructed to make a lexical decision for the last word in each sentence. This experiment allows us to assess the contribution of semantic information to the recognition of upcoming words in reduced speech.

Second, we conducted a word repetition experiment with semantic priming, using the same materials as for the lexical decision experiment discussed above. This time, participants were instructed to repeat the sentence-final word as quickly as possible. By using these two different paradigms, we can investigate the effects of semantic contextual information in the processing of reduced speech from different angles, allowing us to better evaluate the role of semantic information in the processing of conversational speech in everyday listening situations.

Phonetic details as interactional resources

Richard Ogden, Beatrice Szczepek-Reed

Centre for Advanced Studies in Language and Communication, University of York, UK

richard.ogden@york.ac.uk, beatrice.szczepek.reed@york.ac.uk

Much work in sociophonetics broadly speaking seeks to align categories of social identity such as social roles and relationships, or aspects of group identity or the use of a particular style or register with the distribution and use of phonetic features (e.g. many papers in Preston & Niedzielski 2010). In this paper, we take a different view, that of linguistically-informed CA work. Work in this growing tradition explores instead how phonetics provides speakers with resources for social actions. Its primary method of argumentation is to identify how participants themselves orient to phonetic resources (e.g. Couper-Kuhlen & Selting 1996, Szczepek-Reed 2006, 2009, Local & Walker 2005). Social actions studied from this point of view include the management of turn-taking (Wells & Peppé 1996, Ogden 2001, 2004), repair (Selting 1996, Curl 2004), assessing (Uhmann 1996, Ogden 2006) or complaining (Ogden 2010). These works, and many others, argue that speakers use phonetics as one resource among others for the conduct of everyday activity, and do so in principled ways.

In this paper, we will show how segmental phonetic features are used in conversational speech to handle certain kinds of social action. In particular, we will be interested in the projection of action: that is, how a current speaker can mark what a subsequent action might be. This has consequences from the perceptual point of view, because these sorts of detail provide listeners with information about how to interpret what they are hearing, and therefore what next action may or may not be relevant in a subsequent turn. The extract of data below (from Ogden 2001) illustrates the projection phenomenon.

```
Fado 2/26-31
29 C     {C--} {f--}                                    jɑʔ'
       → {öööö} {tai} m:e on oltu     M:adeiralla ja (0.25)
                 or    we  is  be-PPPC Madeira-ADE and
                 or we've been to Madeira and

30               {C}
         paljon  {j}ah
         a lot     and
         a lot and
```

Line 29 contains an inaudibly released glottal stop, followed by a short pause. The recipient does not come in at this place. The unreleased glottal stop is followed in line 30 by a lot and, thus repairing the turn-so-far at line 29. The holding of the glottal stop holds C's turn at least until the talk is repaired. Many other cases of held articulations holding a turn (not necessarily in the conduct of repair) are found in this data. The phonetic detail [held closure] is used in Finnish to project further talk.

Other kinds of segmental detail and their relation to other kinds of social action will be explored in this paper, drawing on data from spoken English and German. Segmental data of this kind is linguistically and psycholinguistically interesting because it falls within a short

time-frame, unlike e.g. long-domain features such as intonation contours (cf. Local 2003). We will argue that in analysing such data, participants' orientation to phonetic detail is an essential tool in making sense of many details of spontaneous speech which seem simultaneously both very obvious and very puzzling; and that many details perhaps loosely ascribable as products of 'spontaneous speech' are in fact used to convey meaning.

Allophonic realizations of the phoneme /t/ in an English spontaneous speech corpus

Benjamin V. Tucker[1], Antti Arppe[1,2]
[1]*University of Alberta, Canada*
[2]*University of Helsinki, Finland*

bvtucker@ualberta.ca, antti.arppe@ualberta.ca

Large corpora of spontaneous speech lend themselves to testing and investigating the realization of various phonemes in the language. In the present study we investigate the allophonic occurrences of the phoneme /t/ as transcribed in the Buckeye Speech Corpus (Pitt et al., 2007). We extracted two forms of context for each instance of /t/ in the corpus: first the dictionary transcription provided by the corpus called the canonical transcription and second the actual realization in the spontaneous speech. We use the contextual environment to investigate the distribution of /t/ allophones in the corpus (following Zue & Laferriere, 1979). We find altogether 29 distinct sounds that are coded as corresponding to the phoneme /t/. Twenty-one of these account for the less than one percent of the total forms. These instances while highly unlikely (e.g. /t/ becomes /m/) and if not coding errors, have generally reasonable alternations when considered with the surrounding context. Overall, the observed range of allophonic outcomes can be seen as gradient from a fully formed obstruent, via various forms of reduction down to outright deletion.

We statistically model the most frequent outcomes using the contextual environments. We use *polytomous logistic regression* according to the *one-vs-rest* technique, which has two key attractive characteristics. First, it models the proportional occurrence of an outcome, such as an allophone, given the occurrence of some combination of linguistic properties (in this case immediately adjacent phonemic context), rather than individual categorical choices. Secondly, the one-vs-rest heuristic highlights the contextual properties which distinguish the individual outcomes (in this case the allophones) from all the rest (within the same set) as odds ratios. We are able to model the distributions of expected probabilities for the outcomes ranging from practically categorical choice, via clear but no absolute preference, to close to equal likelihood given certain left and right contexts for spontaneous speech, supporting the results found in Zue & Laferriere (1979). We also find other statistically likely outcomes not previously reported; importantly the proportion of deletions is relatively frequent. We find that ignoring abstract word boundaries and using actually produced context rather than canonical context is most predictive of the allophonic realizations in the corpus. In other words using the actual spontaneous speech context better models the occurrence of the /t/ allophones than using an abstract lexical input for the context, which result fits with the corpus consisting of continuous speech rather than isolated utterances or words.

Speech timing and accommodation in native and non-native English interactions

Tyler Kendall[1], Ann R. Bradlow[2], Brett Margolis[2]

[1]*Dept. of Linguistics, University of Oregon, USA*
[2]*Dept. of Linguistics, Northwestern University, USA*

tsk@uoregon.edu, abradlow@northwestern.edu

Aspects of speech timing, such as the rate of actual articulation and the distribution of pauses, have been of interest in many areas of language research for many decades. A number of studies have shown that rate and/or pause patterns are influenced by several factors, from cognitive (e.g., Goldman-Eisler 1968, Krivokapic 2007) to demographic (e.g., Kendall 2009, Jacewicz et al. 2009) to interactional (e.g., Staum-Casasanto et al. 2010). Recent work investigating speech rate and pause patterns in a large corpus of sociolinguistic interviews with American English speakers (Kendall 2009, in prep) found that articulation rate differences among talkers were strongly predicted by social factors, such as speaker regional origin, ethnicity, and sex, and, further, that speakers' articulation rates were also somewhat impacted by properties of their interviewers. Silent pauses, however, were found not to be very patterned by these same potential factors and statistical models of the pause data performed quite poorly.

The present project examines articulation rate and silent pause variation in the Wildcat Corpus' collection of native and non-native English speech data (Van Engen et al. 2010). Subjects took part in a *diapix* task, where speakers are paired up to collaboratively solve a spot-the-differences game. This corpus has been used to examine communicative efficiency in spontaneous speech across different language alignment pairs (Van Engen et al. 2010) and provides an excellent resource to look more closely at patterns, and accommodation, in speech timing.

In this paper, we look most closely at the distribution of silent pauses in the spontaneous speech data from the speaker dyads containing different combinations of native English, native Korean, and native Chinese speakers. Unlike Kendall's (2009, in prep) findings, these pause data show strong patterns of accommodation. Native English speakers exhibit the shortest and fewest pauses but also accommodate to the pause patterns of their non-native interlocutors. Native speakers do exhibit faster articulation rates than non-natives but importantly do not appear to accommodate to the non-native speakers in their rates. In this paper, we consider the findings from this study on their own right and in terms of their differences from the American English interview data. We interpret these patterns and differences in terms of speaker- and listener-oriented processes in speech production and, finally, consider whether the differences indicate differences in the types of spontaneous speech elicited in laboratory conditions and in field-based interview settings.

References

Goldman-Eisler, F. (1968). *Psycholinguistics: Experiments in Spontaneous Speech*. London/New York: Academic Press.

Jacewicz, E., Fox, R. A., O'Neill, C., and Salmons, J. (2009). Articulation rate across dialect, age, and gender. *Language Variation and Change* 21(2): 233-256.

Kendall, T. (2009). *Speech Rate, Pause, and Linguistic Variation: An Examination Through the Sociolinguistic Archive and Analysis Project*. Doctoral Dissertation. Durham, NC: Duke University.

Krivokapic, J. (2007). Prosodic planning: Effects of phrasal length and complexity on pause duration. *Journal of Phonetics* 35: 162-179.

Staum Casasanto, L., Jasmin, K., & Casasanto, D. (2010). Virtually accommodating: Speech rate accommodation to a virtual interlocutor. In S. Ohlsson & R. Catrambone (Eds.), *Proceedings of the 32nd Annual Conference of the Cognitive Science Society* (pp. 127-132). Austin, TX: Cognitive Science Society.

Van Engen, K. J., Baese-Berk, M., Baker, R. E., Choi, A., Kim, M. & Bradlow, A. R. (2010). The Wildcat Corpus of Native- and Foreign-Accented English: Communicative efficiency across conversational dyads with varying language alignment profiles. *Language & Speech* 53(4): 510-540.

You don't have to say a word: How duration and F0 trigger or hinder the perception of function words in German

Oliver Niebuhr[1], Evelin Graupe, Laura C. Dilley[2]

[1]*Department of General Linguistics, ISFAS, University of Kiel, Germany*
[2]*Department of Communicative Sciences and Disorders, Michigan State University, USA*

niebuhr@linguistik.uni-kiel.de, evelin_graupe@yahoo.de, ldilley@msu.edu

Much of the focus of spoken language research has shifted from isolated read materials to spontaneous speech. Given the variability of spontaneous speech, how do listeners identify reduced words? Recent work (Dilley & Pitt, 2010; Niebuhr & Kohler, 2011) has shown that listeners interpret durations of speech sounds relative to speech rate context, such that relatively long sounds can be interpreted as cues to reduced or deleted lexical material. The present work aimed to replicate this recent work for German and extend it in two ways. First, we investigated whether it is possible for listeners to perceive multiple lexical interpretations for the same reduced speech. Second, we investigated the contribution of different types of prosodic contexts on which and how many words are perceived.

Experiment 1 materials consisted of sentences containing one of four /n/-initial polysyllabic nouns (*_Nachrichtensprecher*, *_Nebendarsteller*, *No_belpreisträger*, *Na_turheilmittel*); each was preceded by highly-reduced *denn einen*, a particle+article sequence, realized as lengthened noun-initial /n/. Six stimuli were derived from each sentence, based on two basic manipulations: target [n:] was time-compressed or the remainder was time-expanded. Three levels of prosodic manipulation were used: a moderate rate change with original pitch range, a large rate change with original pitch range, or a moderate rate change with expanded pitch range on the accented syllable. Experiment 2 used complementary sentences containing one of the four nouns, but *denn einen* was omitted. In Experiment 2 materials, the /n/ at the beginning of each noun was time-expanded, or the remainder was time-compressed. The same three levels of prosodic manipulation were also used as in Experiment 1. Each experiment additionally included the corresponding original sentences.

Forty-two native German speakers participated in the experiments, 21 in each study; the task was to transcribe each utterance. The frequencies of transcribing *denn einen*, *einen*, *denn*, or Ø prior to the nouns were determined. Mean frequencies revealed that the proportion of polysyllabic responses changed as a function of the relative duration of the target and context in both experiments. Time-compression and time-expansion were both effective at changing lexical percepts. Moreover, expanding the pitch-accent F0 range decreased the amount of perceived lexical material and/or triggered more *denn* transcriptions. The results suggest that listeners formed expectancies about lexical content based on speech rate and pitch-accent context. These expectancies then influenced the number of words and syllables perceived. These findings help to explain the robustness of speech recognition for casual speech.

References:

Dilley, L. C. & Pitt, M. A. (2008). Now you hear it, now you don't: Effects of speech rate on function word perception. Paper presented at the 49 th Annual Meeting of the Psychonomic Society, Chicago.

Niebuhr, O. & Kohler, K. J. (2011). Perception of phonetic detail in the identification of highly reduced words. *Journal of Phonetics 39*, 319-329.

Describing and predicting phonetic reduction in a corpus of spontaneous speech

Philip Dilts, R. Harald Baayen, Benjamin V. Tucker
University of Alberta, Canada

philip.dilts@gmail.com, baayen@ualberta.ca, bvtucker@ualberta.ca

The present investigation describes phonetic reduction in the Buckeye Speech Corpus (Pitt, et al., 2005), and attempts to determine whether and how this phonetic reduction can be predicted. This investigation comprises two studies, one exploring the distribution of phonetic variants of the word-form types in the corpus, and the other investigating possible linguistic influences on reduction in word duration in the corpus. We find that despite large amounts of variation, a statistical model can be produced that achieves a modest amount of success in predicting reduction in word duration.

The first study investigates the prevalence of citation-form pronunciations in the corpus, finding both a surprisingly low prevalence of word-form types that typically appear in such forms (around 60%) and a wide variability between speakers in preference for such forms (from 30%-70% of word types by speaker). The second study uses an underutilized statistical modelling technique, Random Forests (Breiman, 2001) to predict reduction in the duration of words. Random Forests provide a non-parametric modelling technique well suited to high-dimensional data sets with moderately correlated predictors commonly found in linguistics. Measures of interest are entered into a Random Forest regression model, and the model describes the relative importance of each measure in predicting duration reduction. Commonly studied factors such as frequency, conditional probability and rate of speech are all shown to be important predictors of reduction in the Buckeye Corpus. Less thoroughly explored measures like the diversity of pronunciations for a given word type and the predictability of a word's part of speech from surrounding parts of speech are also shown to help predict durational reduction. Surprisingly, age, gender, and interviewer gender were not found to covary with duration reduction.

The first study finds that reduction is common in a corpus of spontaneous speech (see also Johnson, 2004) and that individual speakers vary in the amount of reduction they produce. The second study shows that several linguistic factors can be shown to have a predictable effect on phonetic reduction and that aspects of durational reduction can be statistically modeled.

Between-speaker variability in the clarity of spontaneous speech

Valerie Hazan, Rachel Baker

Department of Speech Hearing and Phonetic Sciences, UCL (University College London), UK

v.hazan@ucl.ac.uk, rachel.baker@ucl.ac.uk

Studies of between-speaker variability in speech intelligibility have shown that individual voices can vary quite significantly in terms of their inherent intelligibility (Bradlow et al., 1996; Hazan and Markham, 2004). These two studies, using read speech materials, highlighted a number of acousticphonetic correlates of intelligibility and found female speakers to be more intelligible than male speakers, on average. A recent corpus of spontaneous dialogs (LUCID corpus) includes orthographically-transcribed speech recordings for 40 young adult speakers (20 women) from a homogeneous accent group while they completed 'spot the difference' picture tasks ('diapix') in pairs (Hazan and Baker, in press). All 40 speakers completed three diapix tasks in one condition where they could hear each other normally (NB condition) and in a condition that involved a transmission channel barrier affecting the other speaker (CB condition).

For each talker, six 2-3 second speech samples were selected per talker per condition; these were excised from around the 10th and 20th turns in the dialog after a number of criteria had been met. These samples were presented to 40 listeners over headphones, randomized across talkers and conditions; listeners rated the clarity of each sample using a 7-point scale (1 very clear – 7 not very clear). For the NB condition, mean clarity ratings varied widely across talkers (range: 2.45 - 4.53). There was a significant effect of talker gender [$F(1,38)=12.8$; $p<0.001$] with a lower (i.e. 'clearer') rating for women (mean: 3.18) than for men (mean: 3.64). Significantly higher clarity ratings were obtained in the CB condition (range: 1.67-3.80; paired-sample $t(39) = -15.4$; $p<0.001$), with again a significant gender effect (women: 2.28; men: 2.72). Clarity ratings were strongly correlated across conditions ($r=0.665$; $p<0.001$).

Global acoustic-phonetic analyses were carried out on the complete speech recordings to obtain measures of mean word duration, vowel F1/F2 range, mean energy and F0 median/range per speaker per condition. On average, these measures were made on between 8 (NB) and 12 minutes (CB) of spontaneous speech per speaker. For both men and women, in the NB condition, only overall F2 vowel range correlated significantly with clarity ratings for speech samples.

In summary, the transmission channel barrier elicited perceptually-clearer speech in the talker not directly experiencing the interference, showing adaptation to the needs of the 'impaired' interlocutor in speech produced with communicative intent. The correlations in clarity ratings across conditions suggest that speakers' ranking in terms of their inherent clarity persists across speaking styles.

References

Bradlow, A. R., Torretta, G. M. and Pisoni, D. B. (1996). Intelligibility of normal speech I: Global and fine-grained acoustic-phonetic talker characteristics. Speech Communication 20, 255-272.

Hazan, V & Baker, R. (in press) Acoustic-phonetic characteristics of speech produced with communicative intent to counter adverse listening conditions. J. Acoust. Soc. Am.

Hazan, V. and Markham, D. (2004). Acoustic-phonetic correlates of talker intelligibility for adults and children. J. Acoust. Soc. Am. 116, 3108-3118.

Tongues in Conversation: a Sociophonetic Comparison of some Scottish Lingual Variables in Conversational and Wordlist Speech

James M Scobbie[1], Eleanor Lawson[1] and Jane Stuart-Smith[2]

[1]Queen Margaret University, Edinburgh, UK
[2]University of Glasgow, UK

Jscobbie@qmu.ac.uk, Elawson@qmu.ac.uk, Jane.Stuart-Smith@glasgow.ac.uk

We report here on conversational data from ECB08, one of the first socially-stratified articulatory corpora. The articulatory instrumentation used was video-based Ultrasound Tongue Imaging (UTI). The majority of the corpus was word-list based, collected from 15 teenage speakers in the Eastern Central Belt of Scotland in 2008. The main socio-articulatory variables examined to date in the wordlist data have been the vowel system, including /u/-fronting (unpublished), showing the extent of /u/ fronting and lowering; and /r/-derhoticisation (Lawson, Scobbie and Stuart-Smith, 2011), showing tongue-shape variation in /r/ as well as in derhoticisation. Social variation in the use of ejective bursts vs. pre-glottalisation in /p t k/ has also been found. In addition, acoustic analysis has been undertaken of some other sociolinguistic variables in a related corpus (WL07), which confirmed that the use of the articulatory instrumentation did not alter speaker behaviour any more than a control group being recorded acoustically (Lawson, Stuart-Smith and Scobbie, 2008).

In this paper we will explore the smaller conversational component of the audio-articulatory corpora, comparing conversational and word-list behaviour.

Speakers were recorded in friendship pairs undertaking some structured discourse-eliciting tasks (e.g. Map Task), followed by a phase of unscripted dialogue. The articulatory conversational data is a sample of the conversational speech. It is not a full record (though a full acoustic recording of both speakers exists): both speakers wore articulatory equipment but only one could be recorded, making a total of eight speakers from ECB08. In addition, the articulatory samples were taken automatically in 15 second contiguous windows, with a gap before the next window of about 10 seconds (while data was saved to disk). Participants were not aware of who was being recorded, or when.

The main variables examined will be /r/ derhoticisation, /u/ fronting, /l/ vocalisation and /t/ glottaling, all of which benefit from articulatory investigation. Initial examination of the data reveals that covert articulations, e.g. from de-rhoticising speakers, are present, particularly in pre-pausal contexts, just as they are in the word-list speech. We will focus on the articulatory similarities and differences between the read-speech corpus of single words, as previously discussed, and the nature of the comparable phenomena in unscripted conversation. We will also look at articulatory movement during silent listening discourse turns. Finally, we will evaluate the conversational corpus, suggesting some methodological improvements.

The phonetics of self-repair in spontaneous Dutch

Paul Carter, Leendert Plug

University of Leeds, UK

p.g.carter@leeds.ac.uk, l.plug@leeds.ac.uk

We present results of a phonetic investigation of instances of self-repair, focusing on the extent to which prosodic factors observed in experimentally-elicited speech are found in a corpus of spontaneous speech.

The psycholinguistic literature contains a number of observations on the phonetics of self-repair. First, Cutler (1983) and Levelt and Cutler (1983) have shown that instances of self-repair in task-oriented dialogue may be prosodically 'marked' or 'unmarked', and that the choice between the two may be constrained by the semantics of the repair – in particular, whether the repair corrects a linguistic or factual error, or refines a correct, but inappropriate formulation. Second, Nooteboom (2005, 2010) has shown that the structural organisation of elicited speech error repairs, in particular whether the word to be corrected or refined is produced completely or not, influences their prosodic realisation. In addition, the interrupted–completed distinction may be associated with the coordination of pre- and post-articulatory self-monitoring processes.

Relatively little has been done to refine Levelt and Cutler's findings, and the phonetic details of 'prosodic marking' in speech repair remain unclear. Moreover, it is unclear to what extent these findings generalise to repairs in genuinely spontaneous speech. The current investigation examines some 500 instances of self-repair sampled from the Spoken Dutch Corpus (Oostdijk, 2002), with a view to modelling their prosodic characteristics.

Preliminary analysis of repairs which correct lexical choices, such as *I'm going by ca- er bike*, and repairs which correct mispronunciations, such as *bark bo- dark boat*, suggests firstly that prosodically 'marked' repairs are relatively scarce in spontaneous speech, and secondly that the explanatory value of the factors mentioned above is limited. The semantics of the repairs (following Levelt and Cutler) have no consistent effect on the pitch, loudness or speech rate of the correct lexical items. The structural organisation of the repairs (following Nooteboom, and also incorporating a more fine-grained temporal approach than the binary interrupted–completed distinction) has some effect mainly on their speech rate characteristics, especially when combined with a simple measure of the phonological complexity of the words involved. These preliminary models only account for a small amount (less than 15%) of the attested prosodic variation, suggesting that the details of speech production in spontaneous speech may differ from the details of speech production in experimentally-elicited speech. We propose that pragmatic factors related to interaction between participants may have a role to play in accounting for the phonetic details.

References

Cutler, A., 1983. Speakers' conceptions of the function of prosody. In: Cutler, A. and Ladd, D.R. (eds), *Prosody: Models and Measurements*. Heidelberg: Springer, 79–91.

Levelt, W.J.M. and Cutler, A. 1983. Prosodic marking in speech repair. *Journal of Semantics* 2, 205–217.

Nooteboom, S. 2005. Listening to one-self: Monitoring speech production. In: Hartsuiker, R.J., Bastiaanse, Y., Postma, A. and Wijnen, F. (eds), *Segmental Encoding and Monitoring in Normal and Pathological Speech*. Hove: Psychology Press, 167–186.

Nooteboom, S. 2010. Monitoring for speech errors has different functions in inner and overt speech. In: Everaert, M., Lentz, T., De Mulder, H. and Nilsen, Ø. (eds), *The Linguistic Enterprise*. Amsterdam: John Benjamins, 213–233.
Oostdijk, N.J.H. 2002. The design of the Spoken Dutch Corpus. In: Peters, P., Collins, P., and Smith,
A. (eds), *New Frontiers of Corpus Research*. Amsterdam: Rodopi, 105–112.

Acoustic duration is predicted by syntactic prototypicality

R. Harald Baayen, Benjamin V. Tucker

University of Alberta, Canada

baayen@ualberta.ca, bvtucker@ualberta.ca

Higher-frequency words tend to have shorter acoustic durations (Bell et al., 2009). According to the smooth signal hypothesis (Aylett & Turk, 2004), speakers adjust the acoustic durations of words such that fluctuations in the rate at which information is transmitted in the speech signal are minimized. In this approach, the effect of frequency on acoustic duration arises as a consequence of speaker-hearer accommodation (cf. Lindblom, 1990). However, it is conceivable that the inverse relation between frequency and acoustic reduction is a consequence of lexical learning. Words that have been learned less well, and that are more difficult to retrieve from lexical memory, would then receive longer acoustic durations.

We report a new predictor co-determining acoustic duration, prepositional relative entropy (PRE). PRE is a measure that captures the extent to which how a given noun (the 'exemplar') makes use of prepositions in simple prepositional phrases such as 'on the table' differs from the average use of prepositions across all nouns ('the prototype') in such phrases. The greater the difference between exemplar and prototype, the greater the PRE is, and the longer response latencies become (Baayen et al., 2011).

The present study extends this result to word naming latencies and fixation durations in reading. Furthermore, words' acoustic durations turned out to be positively correlated with PRE. Importantly, the effect of PRE is orthogonal to that of the word's information load, as captured by minus log frequency. As a consequence, PRE does not contribute to a 'smooth signal', where smooth is defined with respect to the amount of information transmitted per time unit.

Baayen et al. (2011) show that the effect of PRE may arise as a consequence of discriminative learning: Words with prepositional paradigms that are closer to the prototypical use of prepositions are easier to learn, and hence are responded to more quickly in lexical decision and naming tasks. We hypothesize that words that have been learned better can also be articulated more quickly, thanks to higher associative weights on the links from meanings to phonological forms.

Interestingly, the effect of word frequency on acoustic duration can be understood similarly. The informational smoothness of the acoustic signal would then result from the learning of the mapping of meaning and form, obviating the need for teleological explanations in terms of conscious or unconscious speaker-listener accommodation.

Comparison of Reduction in Spontaneous Mandarin and English Speech

L. Ann Burchfield, Ann R. Bradlow

Department of Linguistics, Northwestern University, Evanston, IL, USA

lauraburchfield2014@u.northwestern.edu; abradlow@northwestern.edu

This study aims to compare phonetic reduction in English and Mandarin, with a focus on syllabic reduction. English and Mandarin provide an ideal starting point for cross-language comparison of reduction because of substantial differences in morphology and syllable structure across these two languages. In particular, the restricted syllable structure (no consonant clusters and limited codas) and tight syllable-morpheme correspondence found in Mandarin contrasts with less restrictive phonotactics and looser syllable-morpheme correspondence in English.

Within Mandarin, open syllables are more likely to reduce (e.g. Cheng and Xu, 2009). If this holds true across languages (with the same effect size), we would predict that Mandarin would exhibit more syllabic reduction than English simply due to the preponderance of open syllables in Mandarin. Alternatively, it could be the case that Mandarin's simpler syllable structure means there is less pressure for reduction for the sake of ease of articulation and/or that Mandarin's tight syllable to morpheme correspondence results in greater pressure to keep syllables intact. In this case, we would predict more reduction in English. A final possibility is that all languages are equally redundant in their phonetic encoding of information and can thus afford a similar level of syllabic reduction. Previous work has established some degree of syllabic reduction in English (e.g. Johnson, 2002) and Mandarin (e.g. Cheng and Xu, 2009). The present study directly compares reduction rates by native speakers of English and Mandarin who produced spontaneous narratives in their native language based on the same picture stories. These narratives were transcribed and then read by the same speakers in both clear speech and plain speech. This methodology allows for direct comparison across three styles and across languages with identical materials. For each recording, the number of significant acoustic intensity peaks (acoustic syllables) was automatically extracted (De Jong and Wempe, 2009) and syllable reduction rate was calculated as the ratio of orthographic (text based) to acoustic (signal based) syllables.

In both English and Mandarin, faster speaking rate correlated strongly with increased reduction. The degree of syllabic reduction in the two languages was similar overall, although the pattern across styles varied slightly. These data suggest that the overall rate of syllabic reduction is similar in English and Mandarin despite their differences in morphophonological characteristics. Ongoing analyses will compare the effect of open versus closed syllables on reduction in the two languages and will examine other measures of reduction including contraction of F0 range.

References

Cheng, C. and Xu, Y. (2009) Extreme reductions: Contraction of disyllables into monosyllables in Taiwan Mandarin. In Proceedings of Interspeech 2009, Brighton, UK, pp. 456-459.

De Jong, N.H. and Wempe, T. 2009. Praat script to detect syllable nuclei and measure speech rate automatically. *Behavior research methods* 41(2), 385-390.

Johnson, K. (2002). Massive reduction in conversational English. In Proceedings of the Workshop on Spontaneous Speech: Data and Analysis, Tokyo, Japan.

Acoustic reduction in spontaneous infant-directed speech

Mybeth Lahey[1,2,3] & Mirjam Ernestus[1,2]

[1] *Radboud University, Nijmegen, The Netherlands;*
[2] *Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands;*
[3] *International Max Planck Research School for Language Sciences, Nijmegen, The Netherlands*

mybeth.lahey@mpi.nl; m.ernestus@let.ru.nl

From previous research it is known that speech addressed to young children has a number of properties that distinguish it from speech addressed to adults. One of these properties is that words and segments are pronounced very carefully, leading to the assumption that infant-directed speech is a hyperarticulated register. Spontaneous conversations between adults, on the other hand, typically contain many reduced pronunciation variants. Such reduced variants contain fewer segments or even fewer syllables than canonical pronunciations. This raises the question whether acoustic reduction also occurs in speech to young children.

We investigated reduction in speech addressed to 11- and 12-month-old infants by comparing utterance-medial occurrences of the Dutch words *allemaal* "all" and *helemaal* "completely" from a corpus of conversational infant-directed (52 tokens) and adult-directed (84 tokens) speech and from a corpus of read speech (80 tokens). An acoustic analysis revealed that the tokens in infant-directed speech (mean duration 271 ms) were approximately as long as those in adult-directed speech (mean duration 257 ms), but were generally shorter than the tokens in read speech (mean duration 362 ms). In a rating study, 20 young adult participants listened to all tokens in isolation, rated them for degree of reduction on a six-point scale, and provided a phonetic transcription. The results reflected the acoustic measurements. Listeners perceived approximately the same degree of reduction for tokens in infant-directed (mean rating 4.30) as in adult-directed (mean rating 4.17) speech, but less reduction in read speech (mean rating 1.94). The phonetic transcriptions indicated that participants based their rating scores not only on the duration of the tokens, but also on other properties of the speech signal, including the number of syllables, the number of segments and the presence or absence of segments like the first /l/ and a full vowel in the final syllable. Participants' ratings thus reflected degree of acoustic reduction at several levels.

This study shows that infant-directed speech is not as clearly pronounced as may be expected, and that children are confronted with reduced pronunciation variants from early on in their lives. This suggests that infants acquire full and reduced pronunciations simultaneously, and that for infants these variants may have equal status. Furthermore, our finding that adults did not adapt their use of reduced forms to the linguistic abilities of the infant listeners raises the question whether speakers are able to control their degree of reduction in everyday conversations.

Discourse context and the recognition of reduced and canonical forms

Susanne Brouwer[1], Holger Mitterer[2], Falk Huettig[2,3]

[1]*Northwestern University, Evanston, IL, USA*
[2]*Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands*
[3]*Donders Institute for Brain, Cognition, and Behaviour, Radboud University, Nijmegen, The Netherlands*

smbrouwer@hotmail.com, Holger.Mitterer@mpi.nl, Falk.Huettig@mpi.nl

In two eye-tracking experiments, we examined whether wider discourse information helps the recognition of reduced forms (e.g., ‘puter’) more than the recognition of canonical forms (e.g., ‘computer’). Participants listened to sentences from a casual speech corpus containing canonical and reduced targets while they saw four printed words on the screen: the target (e.g., ‘computer’), a competitor similar to the canonical form (e.g., ‘companion’), a competitor similar to the reduced form (e.g., ‘pupil’), and a phonologically unrelated distractor (e.g., ‘holiday’). Target word recognition was assessed by measuring eye fixation proportions to these printed words.

Experiment 1 presented canonical and reduced forms in a target sentence alone or with an additional discourse context. The additional contexts were samples which directly preceded the target sentences in the casual speech corpus. Results showed that target recognition was facilitated by wider discourse information. Importantly, the recognition of reduced forms improved significantly when preceded by strongly rather than by weakly supportive discourse contexts. Listeners' recognition of canonical forms was, however, not dependent on the degree of supportive context. For canonical forms, the degree of support by wider discourse context allowed participants to predict the target word. But once there was bottom-up information, the degree of support from the discourse context seized to play any detectable role.

Experiment 2 examined whether the effects in Experiment 1 were due to exposure to a speaker’s voice rather than due to discourse information. The same target sentences as in Experiment 1 were presented, but the additional contexts only provided information about the target speaker. These additional contexts were randomly selected samples of the same target speaker. Results showed that the benefits of speaker adaptation were similar for canonical and reduced forms. Moreover, Experiment 2 revealed that the benefits in Experiment 1 are composed of discourse and speaker effects.

In conclusion, our data provide insight into the interplay between prior knowledge and bottom-up information in speech perception. With a clear speech signal, listener can predict upcoming words. However, this prediction appears not to influence word recognition once clear bottom-up information from the target word acoustically unfolds. The situation is reversed for reduced forms. Prediction based on the discourse context is less likely, perhaps due to the poor phonetic quality of the input. Discourse context, however, plays a role in facilitating the recognition of reduced forms. That is, when bottom-up information is unclear, prior knowledge influences word recognition.

Do listeners form expectations about a speaker's tendencies to reduce?

Katja Pöllmann[1,2], Hans Rutger Bosker[3], James M. McQueen[4], Holger Mitterer[1]

[1]Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands
[2]International Max Planck Research School for Language Sciences, Radboud University Nijmegen, The Netherlands
[3]Utrecht Institute for Linguistics OTS, Utrecht University, The Netherlands
[4]Radboud University Nijmegen, The Netherlands

Katja.Poellmann@mpi.nl, h.r.bosker@uu.nl, j.mcqueen@pwo.ru.nl, Holger.Mitterer@mpi.nl

Most research on how reduced forms in casual speech can be recognized centers around the question whether there is lexical storage of reduced variants or whether prelexical processes "undo" the reduction (e.g., Ernestus, 2009). The focus is mostly on the speech signal and on lexical properties. A neglected possibility is that listeners might be able to adapt to the speaking style of their interlocutor. Three eye-tracking experiments examined this possibility. Using a variant of the visual-world paradigm (Mitterer & McQueen, 2009) in Experiment 1 and 2, two groups of participants were exposed to either segmental ([b] > [V]) or syllabic (*ver*-> [f]) reductions. In the test phase, both groups heard both kinds of reductions. Experiment 1 revealed an adaptation effect only for syllabic reductions (i.e., a greater tendency to fixate words with syllabic reductions in the group trained on these reductions). The failure to find an effect for the segmental reduction may have been due to the phonetic implementation of the context of the reduced /b/. Reductions of stops to approximants are articulatory likely after (open) vowels. The /b/ was preceded by a schwa, but this schwa was produced without voicing. Therefore, a second experiment used new materials and a new speaker; the /b/ was now preceded by a full open vowel (/a/), that was realized with voicing. With these materials, Experiment 2 replicated the adaptation effect for syllabic reductions and, importantly, also showed a learning effect for segmental reductions. Experiment 3 investigated the generalization of learning from one type of reduction to another. Three groups of participants were exposed respectively to either segmental ([b] > [m]), syllabic (full vowel deletion) or no reductions. In the test phase, all three groups were tested on both types of reductions. Preliminary results suggest that learning is specific for a given reduction type. Thus, listeners can adapt to a specific reduction style of a speaker, but learning does not generalize from segmental to syllabic reductions or vice versa. Importantly, however, learning about reductions was applied to previously unheard words. This generalization across words suggests that mechanisms compensating for segmental and syllabic reductions take place at a prelexical level.

Frequency Word List of Spontaneous Russian

Irina Apushkina[1], Elena Riekhakaynen[1], Natalia Slepokurova[1], Anatoly Ventsov[2]

[1]Department of General Linguistics, St. Petersburg State University, Russia
[2]Laboratory for Language Behavior Modeling, St. Petersburg State University, Russia

tchiric@mail.ru, reha20@rambler.ru, n.slepokurova@gmail.com, av.ventsov@gmail.com

The word list described in the report has been created in St. Petersburg State University since 2009. Spontaneous-speech dialogues have been collected, segmented into interpausal fragments and given full orthographic description. Selected texts (a radio interview and a part of a TV talk show) of an overall duration of about 50 minutes have been provided with an acoustic-phonetic transcription. Based on the data, a frequency word list has been created for 4682 entries. Every entry is a unique combination of an orthographic description of a word form and its transcription along with the number of occurrences of such a pronunciation in the analyzed texts. In the case of vowel or consonant concatenations at word boundaries (when the separation of two words was very difficult or even impossible) those word forms are described as a single entry.

The set of symbols used in the transcription is partially similar to the X-SAMPA system for computeraided describing phonetic features of a speech signal and consists of Roman alphabet characters with a minimal use of upper case, i.e. capital letters. The transcription is performed by human experts, with the help of digital audio editors. The transcription method enables minimization of the influence of lexico-grammatical experts' knowledge (i.e. auditory analysis and transcription was performed for speech fragments of no longer than onesyllable duration). Even so, instrumental analysis of the word forms taken from spontaneous-speech dialogues has illustrated that experts' transcriptions of various realizations of the same word form can differ significantly while their spectral features are almost completely identical (it mainly concerns vowel identification). In order to minimize the dispersion in experts' decisions a catalogue of typical templates of formant frequencies and trajectories for vowels of different quality has been created and is being used for further transcription.

With the help of the frequency word list the degree of integrity of speech units' segmental structure is being evaluated. About 1/5 of all word-forms diverge from full-type pronunciation, mainly through qualitative or complete quantitative reduction of their elements. The latter and other observations made on the basis of the created word list will be discussed in the report.

When is speech fluent? The relationship between acoustic speech properties and subjective fluency ratings.

Hans Rutger Bosker, Anne-France Pinget, Hugo Quené,
Ted Sanders, Nivja H. de Jong

UiL OTS, Utrecht University, The Netherlands

H.R.Bosker@uu.nl, A.C.H.Pinget@uu.nl, H.Quene@uu.nl, T.J.M.Sanders@uu.nl, n.dejong@uu.nl

The oral fluency level of an L2 speaker is often used as an important measure in assessing language proficiency. In order to improve the objectivity of such language tests, previous studies have attempted to determine the acoustic correlates of fluency (e.g., Cucchiarini et al. 2002). Many of these studies have used multifaceted global measures making the results often difficult to interpret. An example of such a measure is overall speech rate which is confounded because it relates both to speed of articulation and to the use of pauses. Also there is within the literature much diversity in the type of instructions raters were given. Arguing that fluency ratings are dependent on the perception of the acoustic characteristics of speech, Experiment 1 investigated fluency perception by establishing what speech properties raters are capable of perceiving. Three groups of listeners rated the same set of L2 Dutch speech stimuli on either the use of pauses, speed of delivery or the use of repairs (corrections and repetitions). Stimuli were 20sec excerpts from turns in a simulated discussion. Using linear mixed models the subjective ratings were modelled by non-confounded acoustic measures which only measured one aspect of fluency (pause, speed or repairs). Explicit and very specific test instructions resulted in high interrater reliability. Most of the variability of the ratings from the pause group and the speed group was accounted for by pause or speed measures, respectively. Concluding that raters are capable of perceiving and rating pause and speed phenomena (but repair phenomena to a lesser extent), a fourth group of listeners rated the same stimuli on overall fluency. The variability of these ratings was best modelled by pause and speed measures. It is concluded that pause and speed measures are better acoustic correlates of fluency than repair measures. Considering the strong effect of pause measures on fluency perception, Experiment 2 investigates the independent effects of the number of silent pauses and the duration of silent pauses, both in L1 and in L2 speech. Instead of looking at correlations, this experiment attempts to establish a clear causal relationship between these two acoustic speech properties and fluency ratings. By comparing the ratings on identical stimuli differing only in the number or the duration of silent pauses, this experiment reveals whether the number of silent pauses and/or their duration have any effect on fluency perception, both in L1 and in L2 speech.

Cross-linguistic differences in pausing behavior

Nivja De Jong

UiL OTS, Utrecht University, The Netherlands

n.dejong@uu.nl

Pauses in speech can serve communicative means, to help listeners understand (Clark, 1994), and pauses can be due to cognitive factors, when a speaker has not finished planning and formulating the upcoming utterance (Howell & Au-Yeung, 2002). In theories of speech production, lexical concepts are seen as the basic units of planning. If this holds for all languages, one would predict that for an agglutinative language such as Turkish, units of planning can be larger than for a non-agglutinative language such as English. Following this reasoning, speakers of Turkish would have fewer opportunities to pause than speakers of English. This hypothesis is tested by comparing speech data of Turkish and English native speakers. Twenty-four Turkish speakers and twenty-nine English speakers performed eight speaking tasks. These tasks were long turns in simulated conversation. In total, nine hours of Turkish and English speech were annotated, adding information about frequency and duration of silent pauses (as well as other hesitation phenomena).

The results showed that Turkish words are indeed longer in number of syllables and in duration. Furthermore, speakers hardly paused within words, confirming the hypothesis that lexical items form the basis of units-of-speech. Finally, Turkish speakers paused less often than English speakers, but when they paused the duration of these pauses was longer. In total, percentage of time spent pausing did not differ for the Turkish and English speakers. We conclude that usage of pauses due to cognitive factors is dependent on typological features of languages, leading to cross-linguistic differences in pausing behavior.

Lexical hesitation marking in Chintang: Evidence for fillers as words

Tyko Dirksmeyer

Max Planck Institute for Psycholinguistics, The Netherlands
International Max Planck Research School for Language Sciences, The Netherlands

tyko.dirksmeyer@mpi.nl

The status of hesitation markers (or 'fillers', 'filled pauses', 'editing expressions', etc. — such as *uh(m)* in English) has been fiercely disputed in various subdisciplines of the language sciences over the past decades.

Should these items be viewed as aberrations in performance that need to be excluded from linguistic analysis (e.g. Chomsky 1965), are they symptoms of speech production processes that signal trouble but do not signify anything beyond that (Goldman-Eisler 1968; Levelt 1989), or are they actively employed as communicative means just like other words are (Clark and Fox Tree 2002; Jefferson 1974; Schegloff 2010), and thus form an integral part of language?

Chintang, a Tibeto-Burman language spoken in two villages in Nepal, provides evidence for the latter view. Its principal hesitation marker *meĩ* occurs in the same range of functional environments — word search, self-repair, prefacing dispreferred turns, among others — in which *uh(m)* appears in English (and similar forms feature in other wellknown languages). Yet, *meĩ* demonstrably conforms to standard phonological, morphosyntactic and semantic criteria for wordhood, can be seamlessly integrated into utterances, and is regularly exploited for communicative purposes such as "floor management" and projecting what to expect next.

In this talk, I will review data drawn from a corpus of video-recorded naturally-occurring conversational interaction in Chintang and argue for the profoundly conventional nature of hesitation marking with *meĩ*. The findings from this small, as-yet-understudied speech community indicate that fillers should indeed be treated as lexical items on a par with other words. Consequently, they call on linguistic theorizing not only to take hesitation marking and its communicative functions in conversational speech seriously, but also to embrace and incorporate typological diversity in order to arrive at truly generalizable models of language processing.

References

Chomsky, Noam. 1965. *Aspects of the theory of syntax*. Cambridge, MA: MIT Press.

Clark, Herbert H. and Fox Tree, Jean E. 2002. Using *uh* and *uhm* in spontaneous speaking. *Cognition* 84(1):73–111.

Goldman-Eisler, Frieda. 1968. *Psycholinguistics: experiments in spontaneous speech*. New York: Academic Press.

Jefferson, Gail. 1974. Error correction as an interactional resource. *Language in Society* 3(2):181–199.

Levelt, Willem J.M. 1989. *Speaking: from intention to articulation*. Cambridge, MA: MIT Press.

Schegloff, Emanuel A. 2010. Some other "uh(m)"s. *Discourse Processes* 47:130–174.

When words fail: the relationship between gestures and disfluency in child and adult language learners' speech production

Maria Graziano, Marianne Gullberg

The Humanities Lab, Lund University, Sweden

maria.graziano@humlab.lu.se, marianne.gullberg@ling.lu.se

It is often assumed in acquisition and gesture studies that adult and child language learners use gestures as a compensatory device to overcome expressive difficulties, especially representational gestures to either facilitate lexical retrieval, conceptualisation, or information packaging, depending on the theory (Alibali et al. 2000; Kita, 2000; Krauss et al., 2000). This notion is explicit in studies of adult second language learners (e.g., Nicoladis et al., 2007), but is also implicit in studies of child language acquisition (e.g., Goldin-Meadow, 2003; Pine et al., 2007). The theoretical argument hinges on the observation that more difficult tasks, including use of a second language, typically yield higher gesture rates. However, gesture rate alone is not sufficient evidence. It is crucial to examine what types of gestures occur when during difficulties. If gestures are compensatory devices, they should occur in stretches of disfluent speech. However, the evidence is sparse and conflicting for adults, and very little is known about children's speech-gesture production during disfluencies.

This study therefore (a) examines the nature of learners' 'gestural compensation' by comparing gestures produced in non-fluent vs. fluent stretches using a fine-grained temporal analysis of speech and gesture, and (b) compares child and adult language learners' production. The analyses draw on gestures accompanying narrative production by 33 Italian children (4-5; 6-7; 8-10 years) and 16 Dutch adult learners of French as a second language.

Preliminary results indicate that (1) both child and adult learners chiefly produce representational gestures during fluent production, and that any ongoing gesture strokes are interrupted during the onset of a disfluency; (2) gesture strokes that are completed during filled/unfilled pauses are mainly pragmatic gestures indicating an ongoing word search but not its referential content (cf. McNeill, 1985). The rate of such pragmatic gestures increases with age; (3) when speech is resumed, both children and adults also resume gestures.

Learners' gestures are thus interrupted together with speech and they produce different gestures in fluent vs. disfluent speech production. These results have important theoretical implications for both acquisition and gesture research. Moreover, the study provides some of the first evidence that child and adult learners gesture similarly during breakdowns.

A cross-linguistic comparison of speech timing measures in spontaneous and scripted speech: Talker and language specific characteristics in Mandarin-English speakers

Jenna Silver Luque, Laura Ann Burchfield, Kelsey Mok, Ann R. Bradlow

Department of Linguistics, Northwestern University, Evanston, IL, USA

slpjenna@gmail.com, ann@u.northwestern.edu, kelseymok@u.northwestern.edu, abradlow@northwestern.edu

The ALLSSTAR (Archive of L1 and L2 Scripted and Spontaneous Transcripts and Recordings) corpus contains recordings of bilinguals producing scripted and spontaneous speech in both their L1 and L2, as well as English monolingual recordings. This design allows for direct comparison of equivalent materials across and within individual talkers, languages, and speaking styles. In the present study, three groups were compared: native English (n=20), native Mandarin (n=14), and L2 English (n=14) made up of the native Mandarin talkers. By comparing these three groups, we can begin to tease apart language universal characteristics, talker-specific characteristics, and L1 transfer characteristics in both scripted and spontaneous speech.

In this study, we focus on speech timing as reflected in speaking rate and phrase length. All speakers read the North Wind and the Sun passage (IPA Handbook, 1999) and produced a question-prompted 5-minute spontaneous monologue in their native language. Additionally, the native Mandarin speakers carried out these tasks in their L2, English. Each speech recording was segmented into phrases separated by naturally-occurring silences of 100ms or more. A Praat script (DeJong and Wempe, 2009) was used to extract the number of significant intensity peaks (acoustic syllables). Phrase length and speaking rate were calculated as the number of acoustic syllables per phrase (aspp) and acoustic syllables per second (asps), respectively.

In the scripted speech, L1 Mandarin showed significantly shorter phrases (8.3 aspp) than L1 English (16.2 aspp), while L2 English showed a phrase length similar to that of L1 Mandarin (9.2 aspp). The difference across L1s could represent a difference in phrase level organization for reading in English and Mandarin, with L2 English showing L1 transfer. However, in the spontaneous speech, the two L1s had phrases of similar average length (English= 9.6, Mandarin=9.4 aspp), but L2 English phrases were markedly shorter (6.0 aspp). In both scripted and spontaneous speech L2 English showed a slower speaking rate (scripted=3.1 asps; spontaneous=2.8 asps) than L1 English (scripted=3.7 asps; spontaneous=3.2 asps) or L1 Mandarin (scripted=3.5 asps; spontaneous=3.5 asps), replicating the well-known rate decrease for L2 speech. Importantly, within the native Mandarin speakers, we found strongly positive correlations between L1 and L2 phrase length ($r=0.69$, $p<.02$) and speaking rate ($r=0.78$, $p<.005$) for spontaneous speech but not for scripted speech. Together, these findings suggest that some aspects of speech timing in spontaneous but not scripted speech may reflect talker-specific patterns that transcend the native or non-native status of the language being spoken.

References

De Jong, N.H. and Wempe, T. 2009. Praat script to detect syllable nuclei and measure speech rate automatically. Behavior research methods 41(2), 385-390.

The Handbook of the International Phonetic Association. 1999. Cambridge University Press.

"Sorry, what was that?": The roles of pitch, duration, and amplitude in the perception of reduced speech

Ryan Podlubny, Benjamin V. Tucker, Terrance M. Nearey

University of Alberta

pudplace@gmail.com, bvtucker@ualberta.ca, t.nearey@ualberta.ca

Reduced forms in spontaneous speech often bear little resemblance to their carefully produced counterparts. Context (phonetic, syntactic and semantic) plays an important role in processing these reductions (Ernestus et al., 2002) and often succeeds to the degree that many reductions are hardly noticed by the listener. The present study considers the level of reduction, the context in which it occurs, and the importance of specific phonetic information available in reduced productions. Combined, these factors contribute to a listener's ability to extract meaning from reduced forms; and by manipulating specific acoustic elements we can clarify the conditions under which listeners find speech intelligible. In this project we focus on the importance of pitch, amplitude and duration in reduced speech under conditions of partially or totally 'whitened' local spectral properties.

A range of reduced forms (judged impressionistically) were extracted from a spontaneous speech database taken from a female speaker, representative of Western Canadian English. Target items were taken with phrase-level context to be used in three experiments - each designed to test the degree of contribution from a particular aspect of the signal corresponding to a potential phonetic cue. In the first experiment, targets were altered by producing several different signal-to-noise ratios (SNR) using signal correlated noise (SCN), which preserves intensity of the original speech at all time scales. Contextual frames were not manipulated in this study. With pure SCN (- infinity SNR) amplitude envelope information is preserved, while information from pitch and short-term spectrum envelope are eliminated. Altered and unaltered tokens were played, randomly, through headphones to listeners who were instructed to type out what they believed the speaker said. These orthographic transcriptions were scored against a master transcription agreed upon by the experimenters. Pilot results indicate that reduced speech is not perceptually restored with very low SNR with SCN, while in better SNRs listeners recover considerably more information. This suggests that intensity alone does not offer enough information for phonetic restoration of reduced speech, and that some aspects of pitch and short-time-spectrum are important. Experiment 2 will focus on preserving pitch and intensity information, while bleaching the spectral envelope by using an intensity-scaled (to match the original signal) LPC residual. Experiment 3 will focus on global duration, lengthening or shortening the target signals using pitch-synchronous overlap add or LPC methods. Our results will help identify the contribution of phonetic cues in the perception of reduced speech.

References:

Ernestus, M., Baayen, H., and Schreuder, R. (2002) The recognition of reduced word forms. *Brain and Language* 81, pp.162-173

Coverage of Spontaneous Conversational Speech from Nijmegen Corpus of Casual Czech by General ASR Language Models

Vaclav Prochazka, Petr Pollak

Faculty of Electrical Engineering, Czech Technical University in Prague, Czech Republic

prochva1@fel.cvut.cz, pollak@fel.cvut.cz

The Large Vocabulary Continuous Speech Recognition (LVCSR) as one of the frequent applications of speech technology is being applied nowadays in growing number of applications in everyday human life. Consequently, also the need of spontaneous speech recognition arises, however, such speech has strongly different character in comparison to non-spontaneous speech. Then such specific phenomena are not supposed to be covered by standard general Language Model (LM).

In this contribution we will analyze Nijmegen Corpus of Causal Czech (NCCCz) collected under our co-operation with Radboud University of Nijmegen. We will analyze the content of this corpus from the point of view of several LMs which are publicly available, i.e.:

1) Czech LC-Star lexicon created within European project LC-StarII,
2) LMs created from Czech National Corpus (CNC),
3) LMs from publicly available WEB1T 5-gram corpus (WEB-based corpus).

Within this study we will analyze the following criteria:

- the rate of Out-Of-Vocabulary (OOV) words,
- the rate of word fractions, word repetitions, or repeated starts as typical phenomena for spontaneous speech,
- the perplexity computed at text level above transcription of Nijmegen corpus of casual speech,
- speech recognition performance above recordings from NCCCz using above mentioned language models,
- finally, the first attempts with adaptation of LMs for better description of spontaneous speech nature will be presented.

The following table summarizes the results of OOVs computed for different corpora and lexica.

	all NCCCz transcriptions:			NCCCz transcriptions without word-fractions:			the corpus of texts for the testing of LVCSR:		
lexicon	lex-size	words	oovrate	lex-size	words	oovrate	lex-size	words	oovrate
cnc	60000	371011	10.50	60000	362303	8.35	60000	32185	5.18
cnc	60000	371011	5.68	340000	362303	3.41	340000	32185	1.48
cnc	861899	371011	4.95	861899	362303	2.67	861899	32185	1.14
web	60000	371011	9.09	60000	362303	6.91	60000	32185	7.24
web	340000	371011	5.26	340000	362303	2.98	340000	32185	1.60
web	957285	371011	4.83	957285	362303	2.54	957285	32185	1.13
lcstar	84724	371011	8.88	84724	362303	6.69	84724	32185	5.78

These results proved that WEB1T LM contains more words from spontaneous speech as large WEB-based collection should contain also texts from chats which can contain some tokens similar to spontaneous speech. Also OOV rates decreased when word fractions were removed from NCCCz corpus but not so much to be comparable with reference non-spontaneous corpus used e.g. for LVCSR testing.

Filled Pauses in Writing: What can they Teach us about Speech?

Ralph Rose

Faculty of Science and Engineering. Waseda University, Tokyo, Japan

rose@waseda.jp

This presentation reports on a research effort to use filled pauses ('uh', 'um': hereafter, FPs) in blog writings to better understand how and why speakers use them in spontaneous speech. Blog FPs are written intentionally and cannot be the result of some linguistic processing shortcoming (i.e., speech-repair as in Levelt, 1983). Hence, if written FPs can be accurately characterized, then the spoken FPs that fit this characterization can be removed from consideration leaving a smaller, potentially more uniform set of other FPs for further study.

Samples of FPs in blog writings were gathered from 100 top blogs. Samples of FPs in spontaneous speech were taken from the Switchboard corpus. A balanced sample of 227 FPs were gathered of each type. Each FP was categorized according to its medium (written or spoken), its location (at clause boundary or clause-internal), the part-of-speech of the immediately following word (content or function, following Maclay and Osgood's 1959 classification), and the FP type (open 'uh' or closed 'um', after Rose, 1998). The data was analyzed under a generalized linear model with chi-square tests.

There was a main effect of FP Type (Chi-square=48.4, $p<0.001$) with a ratio of open to closed FPs of approximately 2:1. This is comparable to previous studies (e.g., Rose, 1998). There were no other main effects. There was an interaction between medium and following word type (Chi-square=37.0, $p<0.001$), as well as between medium and FP type (Chi-square=5.4, $p<0.05$). In the spoken medium, the following word was 30% more likely to be a function word than a content word, while in the written medium, this trend reversed: the following word was 70% more likely to be a content word than a function word. Also, in the spoken medium, the ratio of open to closed FPs was almost 3:1, but in the written medium, this ratio dropped to 1.4:1.

Results from FPs in writing suggest a hybrid view of FPs in speech: Some FPs are used intentionally and with some selectional restrictions (i.e., before content words) in order to serve some pragmatic function (cf., filler-as-word hypothesis in Clark and Fox Tree, 2002), with open FPs being slightly preferred in this role. Other FPs in speech are the result of difficulties during linguistic processing and occur semi-automatically as part of speech repair (cf., Levelt, 1983).

References

Clark, H., & Fox Tree, J. E., 2002. Using uh and um in spontaneous speaking. Cognition, 84:73–111.

Levelt, W.J.M., 1983. Monitoring and self-repair in speech. Cognition, 14: 41-104.

Maclay, H. and Osgood, C.E., 1959. Hesitation phenomena in spontaneous English speech. Word,15: 19-44.

Rose, R.L., 1998. The communicative value of filled pauses in spontaneous speech. Master's Dissertation, University of Birmingham, UK.

Reduced Word Forms in the Mental Lexicon: Evidence from Russian

Elena Riekhakaynen, Olga Raeva

Department of General Linguistics, St. Petersburg State University, Russia

reha20@rambler.ru, olgaspace@rambler.ru

The fact that the interpretation of reduced word forms, which are quite frequent and diverse in casual speech, does not normally cause any difficulties for a listener makes it plausible that all phonetic variants of a word form are stored in the mental lexicon and accessed directly in the process of spoken word recognition.

However, the results of an experiment in which subjects were listening to isolated reduced word forms extracted from spontaneous dialogues in Russian and the same word forms in context refute this hypothesis. As well as in the experiment on Dutch described in (Ernestus et al., 2002), the percentage of misidentifications of highly reduced Russian forms in isolation was quite high, whereas in the context these forms were well recognized. It makes us agree with the assumption expressed in article mentioned above that only unreduced forms are stored in the mental lexicon, all other variants being reconstructed using contextual information and some phonetic cues. According to the results of our experiments, consonants (especially consonant order and single initial stop consonant) are more reliable sources of information for a listener than vowels in the process of Russian reduced word forms recognition.

Words that normally appear in a highly reduced form in casual and even prepared speech seem to be the only exception to the rule, their most typical variant being recognized accurately even in isolation or limited context (e.g. Russian *сейчас* '*now*' /s'ičás/ that is normally pronounced as /š':as/, etc). Presumably, these typical realizations can be stored in the mental lexicon along with unreduced variants, especially if reduction affects the consonant structure making the process of reconstruction to the unreduced form more complicated. Analysis of 20 highest-frequency word forms taken from spontaneous Russian speech of an overall duration of
about 260 minutes has shown that such forms are not numerous in Russian. The majority of analyzed word forms are produced mostly in non-reduced manner, and among the reduced ones prevail those with one-sound (most often — vowel) elision, which can be easily 'reconstructed' when compared to the respective fully pronounced forms being kept in the mental lexicon.

References

Ernestus, M., Baayen, H., & Schreuder, R. (2002). The Recognition of Reduced Word Forms. *Brain and Language*, 81, 162–173.

Reduction of the discourse marker *you know*: production and comprehension

Louise Schubotz[1], Mirjam Ernestus[1,2], Nelleke Oostdijk[1]

[1]*Radboud University Nijmegen, The Netherlands*
[2]*Max Planck Institute for Psycholinguistics, The Netherlands*

louise_schubotz@gmx.de, m.ernestus@let.ru.nl, n.oostdijk@let.ru.nl

In spontaneous conversations, phrases like *you know* may be used as discourse markers. Previous research strongly suggests that these phrases tend to be acoustically more reduced if used as discourse markers than with their literal meanings. The present study contributes to this research by investigating how pragmatic function correlates with several acoustic characteristics of the phrase *you know* in a large corpus of American English and by investigating which of these characteristics help listeners identify this phrase as a discourse marker.

We extracted 300 tokens of *you know* from the Buckeye corpus of conversational speech, 63 of which could be classified as tokens used with their literal meaning while 237 classified as discourse markers. Statistical analyses showed that the duration of *you know* is influenced by the local speech rate and the presence of pauses, especially if the phrase is used with its literal meaning. Apparently, the discourse marker is prosodically more independent from its context. This is in line with our informal observation that the discourse marker is also more often preceded by a pause and seems to form an intonation contour of its own. Moreover, in the discourse marker the vowel of *you* and the /n/ of *know* are reduced more often than is the case with the literal use of the phrase. These results show systematic differences in the realization of *you know* used as a discourse marker and with its literal meaning.

In a perception experiment we investigated which acoustic characteristics help listener identify a token of *you know* as a discourse marker. We selected 46 tokens (18 used with the literal meaning and 28 discourse markers) of which the pragmatic function is not revealed by the preceding semantic/syntactic context. Twenty-five native speakers of Dutch, highly proficient in English, scored the tokens as discourse marker or as used with their literal meaning. They read the preceding context including *you know* for half of the tokens and heard the corresponding speech signal for the other half. The speech signal was manipulated such that pauses preceding *you know* were deleted and the intonation contour was flat. Participants correctly identified a token as a discourse marker more often in the auditory (81%) than in the visual condition (50%). They tended to identify a token as used with its literal meaning if it contained a full vowel in *you* and the /n/ of *know* was clearly audible.

These results show that phrases may be more reduced if used as a discourse marker and that listeners can use this acoustic information to interpret the pragmatic function of these phrases.

The social perception of turn-taking cues in spontaneous conversation

Marisa Tice[1], Tania Henetz[2]
[1]*Linguistics, Stanford University, USA*
[2]*Psychology, Stanford University, USA*

middyp@stanford.edu, thenetz@stanford.edu

Turn-taking necessarily involves precise timing and interactivity; it requires complex cognitive and social processes that speakers must perform online while hearing (and sometimes producing) spontaneous speech. In conversation timing is key, and speakers monitor turn-taking cues on-line to continuously update their expectations of their interlocutors. This study investigates whether inter-speaker gap duration, overlap, and backchanneling play a role in the social perception of speech during conversation.

Recent work shows that listeners attend to turn-taking cues as indications of the speaker's knowledge level (Roberts et al., 2006). Further, Pearson et al. (2008) found that the social perceptions of response-timing interact with race when interracial dyads reported greater anxiety and less interest in conversation after a 1-second timing lag was (covertly) introduced to their video chat.

In this study, participants heard clips of spontaneous, dyadic conversation in a matched-guise task. Clips were phonetically-manipulated to control for the three turn-taking cues of interest (inter-speaker gap duration, overlap, and backchanneling) by stretching and shrinking gap durations and replacing the naturally-occurring backchanneling with silence. Participants then rated the speakers and conversations on a range of social measures, including dominance, formality, and similarity.

Initial results indicate that listeners' social judgments are sensitive to these cues. For example, the removal of backchanneling in a male-male conversation resulted in ratings of higher speaker interest and lower speaker similarity, running contrary to some previous descriptions of backchannels as supportive signals (Yngve, 1970). This effect may interact with gender, such that backchanneling is interpreted differently in female-female or mixed-gender dyads, supporting Pearson et al.'s (2008) finding that turn-timing effects interact with (mis)matches in speaker race. It is clear these judgments follow from a complex interaction of subtle acoustic cues, social information, and context that interlocutors evaluate on-line in conversation.

Studying speech styles from authentic data: dramatic prosody in live football commentaries

Jürgen Trouvain[1], Friederike Kern[2]

[1]*Saarland University, Germany*
[2]*Hildesheim University, Germany*

trouvain@coli.uni-saarland.de, kernfr@uni-hildesheim.de

Although there are many advantages of "scripted lab speech" as well as "non-scripted lab speech" the recordings are made deliberately for the sake of empirical research, not for communication. One alternative to this authenticity problem is to use naturally occurring data produced by professional speakers, with real speaking roles and conversational tasks deriving from them. TV and radio broadcasted speech meets these demands as the reporters are usually trained speakers keen to solve the conversational tasks at hand. For the study presented here, we chose sports commentaries featuring suspense and/or drama by extreme prosodic means, as our research will show.

According to previous research, radio football commentaries exhibit different speech styles according to the various conversational tasks [1]. Special attention has been paid to the signalling of suspense that is one of the core tasks of the radio reporters. It is systematically achieved by a speech style called "speaking dramatically" with prominent prosodic characteristics which are similar to those found for horse race commentaries [2]. In football commentaries, the climax of the suspense-building phase can be reflected by a "goal roar" with extremely high pitch; however, differences could be found between commented goals for or against the own team, and between primarily listener-oriented (radio) commentaries or television commentaries [3].

For the present study, one football match (England v. Germany Men's World Cup quarter final 2010) with four different commentaries was selected: from German public television (only one commentator), German public and private radio and public English radio (two speakers respectively). The six goal scenes were marked by a continuous rising of pitch in all commentaries (often exceeding two octaves). Most German radio commentators articulate very fast while describing a goal potential scene but slow down considerably when describing the goal itself. Their English colleagues, on the contrary, often keep a rather constant articulation rate all along. Further features of speaking dramatically include an overriding of pitch accents and boundary tones and a sudden change in voice quality.

These results show the special characteristics of "speaking dramatically" within the genre of sports commentaries. The analysed non-lab data strongly contrast to data taken from lab speech styles. Obviously, real communication situations provide far more and far extreme phonetic variation than speech recorded in the lab was able to produce so far.

References

[1] Kern, F. 2010. Speaking dramatically: The prosody in radio live commentaries of football games. In: Selting, M., Barth-Weingarten, D. & Reber, E. (eds): Prosody in Interaction. Amsterdam: Benjamins, 217-238.

[2] Trouvain, J., & Barry, W.J. 2000. The prosody of excitement in horse race commentaries. Proceedings of the ISCA-Workshop on "Speech and Emotion", Newcastle (N. Ireland), 86-91.

[3] Trouvain J. 2011. Between excitement and triumph – live football commentaries in radio vs. TV. In: Proceedings of the 17th International Congress of Phonetic Sciences, Hong Kong.

Sociolinguistic Expectations and the Segmentation of Conversational Speech

Kodi Weatherholtz

The Ohio State University, USA

kweatherholtz@gmail.com

Previous research has shown that activation of social categories robustly influences speech perception (Strand, 1999; Hay & Drager, 2010). Currently though, little is known about effects of social categories on processing of longer, more conversational utterances (however, see Staum Casasanto, 2008). Here we present results from two experiments testing whether sociolinguistic expectations influence the parsing of juncture ambiguities such as [rɪ.tʌr.nɪn.], which in rapid speech can signal either the polymorphemic word *returnin'* or a phonetically reduced *return in*.

Realization of the morpheme (ING) varies by race and city-country orientation, *inter alia*, with white and city-oriented males using fewer alveolar (ING) forms like *returnin'* than black and country-oriented males respectively (see Hazen, 2005). By contrast, use of verb-preposition sequences has not been linked to social category variation. Thus, we predict that listeners can exploit probabilistic knowledge about the social distribution of alveolar (ING) use to anticipate whether a production like [rɪ.tʌr.nɪn.] is a speaker's *returnin'* or *return in*.

In two experiments, listeners heard locally ambiguous sentences as in (1) that are disambiguated sentence-finally. Target utterances were spoken by two Caucasian male speakers, and a picture of an ostensible speaker was presented along with each utterance to manipulate this speaker's race (experiment 1) and city-country orientation (experiment 2). Response times were recorded as listeners determined whether targets "made sense".

(1) a. We saw the man *returnin'* a crappy gift. V+(ING) sequence
b. We saw the man *return in* a crappy mood. V Prep sequence

The critical prediction is that listeners will respond slower after receiving the disambiguating noun *gift*, which requires segmentation as V+(ING), when the speaker is a white or city-oriented male then when the speaker is a black or country-oriented male since the former are relatively less likely to produce alveolar (ING). Such would indicate that listeners use sociolinguistic expectations to segment running speech into words.

Experiment 1 partially confirmed these predictions. For one speaker, listeners responded significantly slower to the V+(ING) tokens when the speaker was presented as white than as black. However, for the second speaker, no effect of race was found. Unexpectedly, voicepicture pairings in some cases were no believable. Additionally, utterances were not perceived as equally ambiguous.

Experiment 2 is currently underway using a similar design but with a city-country manipulation. To address aforementioned complications, new speakers were recorded, utterances were piloted to locate truly ambiguous ones, and voice-picture pairings were piloted to ensure believability.

References

Hay, J. and Drager, K. (2010) Stuffed toys and speech perception. Linguistics, 48(4):865-892.
Hazen, K. 2005.The in/ing variable. In K. Brown, editor, Encyclopedia of Language and Linguistics, volume 5. Elsevier, 2 edition.
Staum Casasanto, L. 2008. Experimental Investigations of Sociolinguistic Knowledge. Dissertation, Stanford.
Strand, E. 1999. Uncovering the role of gender stereotypes in speech perception. Journal of Language and Social Psychology, 18(1), 86-99.

F0 Declination in French : Broadcast News versus spontaneous speech

Cedric Gendrot, Martine Adda-Decker, Carolin Schmid

Laboratoire de Phonétique et Phonologie, Université Paris Sorbonne Nouvelle, France

cgendrot@univ-paris3.fr, madda@limsi.fr, schm2801@uni-trier.de

This study compares F0 declination in French from two kinds of corpora : Broadcast News speech and spontaneous speech. F0 Declination refers to the downward trend of F0 over the course of an utterance. It has been found that declination is expected and used by listeners. However, it is unclear whether declination is part of the linguistic code and speaker-controlled, or whether it is an automatic byproduct of some physiological process (tracheal pull, the downtrend of subglottal pressure, and/or the activity of laryngeal muscles). The nature of declination is also debated: is it a global attribute that requires whole-phrase planning, or is it a concatenation of local events?

Contrary to most previous studies of declination, we investigate F0 declination in large uncontrolled corpora (around 30 hours for each corpus, i.e. 60 hours altogether). We expect that by using large and natural speech corpora, most factors influencing F0 declination will balance out, thus revealing the general declination effect that we are interested in. We will compare the declination effect between Broadcast News Speech (taken from the ESTER corpus) and spontaneous speech (taken from the Nimejen speech corpus). Such a comparison should help us better understand the nature of declination.

As previously done by Yuan and Liberman, we extracted the “utterances” that are time-stamped in the human transcripts. Moreover, as a confirmation procedure, we also used pauses (different durations are tested) detected by the LIMSI automatic alignment in order to select the “utterances”. Utterances containing a pause longer than 50 ms were excluded.
The final data include 9,800 broadcast utterances and 7,760 spontaneous utterances. The F0 contours of the utterances were extracted using PRAAT with a 10 ms frame rate. The contours were then linearly interpolated to be continuous over the unvoiced segments, smoothed and then converted to semitones. Two methods were applied to measure F0 declination. First, a linear regression line was fitted to each F0 contour and the slopes of the fitted lines were then analyzed. Secondly, we used a convex-hull algorithm to identify local F0 valleys and peaks and thus obtain a top and a bottom line.

Analysis of the data demonstrated a strong correlation between declination slope and utterance length: the shorter the utterance, the steeper the declination. Both the topline and baseline show declination, and the topline has final lowering in both corpora. In spontaneous speech, the baseline is close to a straight line, which is different from its topline. In broadcast speech however, the baseline and topline are similar, both consisting of three parts: initial rising, middle declination, and final lowering. The result suggest that the declination slope is controlled by speakers, and that there is preplanning on declination in speech production.

Interlocutor accommodation across communication barriers in naturalistic conversations

Ann R. Bradlow[1], Valerie Hazan[2], Midam Kim[1], Michele Pettinato[2]

[1]*Department of Linguistics, Northwestern University, Evanston, IL, USA*
[2]*Department of Speech Hearing and Phonetic Sciences, University College London, UK*

abradlow@northwestern.edu, v.hazan@ucl.ac.uk, midamkim@gmail.com, m.pettinato@ucl.ac.uk

Real-world conversational speech is typically a joint rather than individual activity with many opportunities for interlocutor accommodation. Previous work has demonstrated listener-oriented speech adjustments including foreigner- and child- directed speech (e.g. Uther, Knoll, Burnham, 2007) and speech directed to a hearing-impaired listener or in noise (e.g. Picheny, Durlach, Braida, 1985). Moreover, the presence of a real rather than hypothetical interlocutor influences speech production (Charles-Luce, 1997; Scarborough et al., 2007), suggesting that listener-directed modifications are best characterized under conditions of authentic communicative intent.

The present work examined communicative efficiency and acoustic-phonetic enhancement in dialogues elicited during a cooperative picture-matching ("diapix") task under conditions of either a transmission channel barrier or linguistic barrier. Efficiency was assessed by task completion time (TCT) and word type-to-token ratio (TTR). Acoustic measurements included F0 median and range, mean energy in the mid-frequency range and average word duration. Data came from two English diapix corpora, LUCID (Hazan & Baker, In press) and Wildcat (Van Engen et al., 2010); see table for conditions in both corpora. For the LUCID transmission channel conditions, only speech by the talker hearing normally was analyzed. Despite substantial differences in design, equipment, location and diapix scenes, we applied consistent measures to both corpora to compare the effects of different barrier types.

	'Control' condition	Transmission channel barrier	Linguistic barrier
LUCID	1. NB (20 pairs)	1. VOC (20 pairs) 2. BAB (10 pairs)	1. N-NN (10 pairs)
Wildcat	1. N-N (8 pairs)		1. N-NN (8 pairs) 2. NN1-NN1 (11 pairs) 3. NN1-NN2 (10 pairs)

NB = No barrier, N=Native, NN=Nonnative, VOC = Vocoded speech transmission, BAB = Multitalker babble background, NN1-NN1 = nonnatives with matched L1, NN1-NN2 = nonnatives with mismatched L1.

Results showed significant decreases in efficiency (increased TCT, decreased TTR) for both transmission channel (LUCID: VOC, BAB) and linguistic barriers (LUCID N-NN, Wildcat N-NN, NN1-NN1, NN1-NN2) relative to control conditions (LUCID NB, Wildcat N-N) with a greater effect size for the linguistic conditions (involving two-way communication difficulties) than the transmission channel barriers (involving one-way communication difficulties). While there were consistent and reliable acoustic enhancements in the within-talker LUCID comparisons between the NB and communication barrier conditions, the between-talker Wildcat comparisons showed variable acoustic enhancements. Within LUCID, the acoustic enhancements were less extensive with a linguistic barrier than a

transmission channel barrier. Taken together, these findings support the view that speech accommodation in naturalistic dialogues varies depending on the nature of the communication channel.

References

Charles-Luce, J. (1997).Cognitive factors involved in preserving a phonemic contrast. *Language & Speech*, 40(3):229-48.

Hazan, V & Baker, R. (In press) Acoustic-phonetic characteristics of speech produced with communicative intent to counter adverse listening conditions. *Journal of the Acoustical Society of America.*

Picheny, M., Durlach, N., & Braida, L. (1985). Speaking clearly for the hard of hearing I: Intelligibility differences between clear and conversational speech. *Journal of Speech and Hearing Research*, 28, 96-103.

Scarborough, R., Brenier, J., Zhao, Y., Hall-Lew, L., & Dmitrieva, O. (2007) An Acoustic Study of Real and Imagined Foreigner-Directed Speech. *Proceedings of the International Congress of Phonetic Sciences*, 2007.

Uther, M., Knoll, M. & Burnham, D. (2007). Do you speak E-NG-L-I-SH? A comparison of foreigner- and infant- directed speech. *Speech Communication*, 49, 2–7.

Van Engen, K. J., Baese-Berk, M., Baker, R. E., Choi, A., Kim, M. & Bradlow, A. R. (2010). The Wildcat Corpus of Native- and Foreign-Accented English: Communicative efficiency across conversational dyads with varying language alignment profiles. *Language & Speech*, 53(4), 510-540.