



The Buckeye corpus of conversational speech: labeling conventions and a test of transcriber reliability

Mark A. Pitt ^{a,*}, Keith Johnson ^b, Elizabeth Hume ^b, Scott Kiesling ^c,
William Raymond ^b

^a *Department of Psychology, Ohio State University, 1885 Neil Avenue, Columbus, OH 43210-1222, USA*

^b *Department of Linguistics, Ohio State University, Columbus, OH 43210, USA*

^c *Department of Linguistics, University of Pittsburgh, Pittsburgh, PA 15260, USA*

Received 16 January 2004; accepted 23 September 2004

Abstract

This paper describes the Buckeye corpus of spontaneous American English speech, a 307,000-word corpus containing the speech of 40 talkers from central Ohio, USA. The method used to elicit and record the speech is described, followed by a description of the protocol that was developed to phonemically label what talkers said. The results of a test of labeling consistency are then presented. The corpus will be made available to the scientific community when labeling is completed.

© 2004 Elsevier B.V. All rights reserved.

Keywords: Spontaneous speech corpus; Transcription; Labeling; American English

1. Introduction

The purpose of this paper is to introduce the Buckeye corpus of conversational speech. After explaining why it is being created, we describe

how the speech was collected, a few characteristics of the corpus, the procedure developed for phonemic labeling, and a test of transcription consistency.

2. Why create the corpus?

Creation of the Buckeye corpus was born out of an interest in phonological variation and its effects on speech recognition by humans and machines. Phonological variation occurs naturally in speech

* Corresponding author. Tel.: +1 614 292 4193; fax: +1 614 688 3984.

E-mail addresses: pitt.2@osu.edu (M.A. Pitt), kjohnson@julius.ling.ohio-state.edu (K. Johnson), ehume@julius.ling.ohio-state.edu (E. Hume), kiesling@pitt.edu (S. Kiesling), raymond@ling.ohio-stat.edu (W. Raymond).

production, no matter whether the talker is speaking slowly or rapidly, and varies within as well as between talkers. It can result in nontrivial distortions in the realization of a word because the changes that occur can be perceptually meaningful distinctions in the language, not simply random acoustic variation. For example, the word *went* can be pronounced as [wɛnt], [wɛn] and [wɛnt], among others, with the last two forms illustrating the deletion and substitution of a single phone. In the case of [wɛn], deletion of [t] creates a potential confusion with the word *when*. When spoken with the reduced vowel [ə], [wɛnt] could be misheard as *want*.

Phonological variation is of considerable interest to researchers throughout the speech sciences. Linguists have amassed collections of spoken variants and categorized them into different types of variation (e.g., deletions, substitutions, insertions) and as a function of the surrounding contexts that promote and inhibit their manifestation (Brown, 1990; Gimson, 1989). An additional aim of this work has been to estimate the frequency and regularity (i.e., lawfulness) of particular types of variation (Dalby, 1986; Greenberg, 1997; Guy, 1980; Jurafsky et al., 1998; Labov, 1994; Neu, 1980; Shockey, 1973).

The availability of phonemically labeled speech corpora such as TIMIT and Switchboard has prompted linguists to study the acoustic/phonetics of phonological variation with the aim of providing a richer and deeper understanding of the phenomenon of variation (Byrd, 1992, 1993, 1994; Keating et al., 1994; Jurafsky et al., 2001; Manuel et al., 1992). Corpus-based investigations like these are attractive because one is able to study efficiently phonetic variability in connected speech over a large number of talkers.

Phonological variation is also of interest to researchers at the other end of the speech chain: speech recognition, be it by humans or machines. Variation is a nontrivial problem for scientists studying automatic speech recognition (ASR). In their review article “Modeling pronunciation variation for ASR: a survey of the literature”, Strik and Cucchiaroni (1999) state that “pronunciation variation research in ASR still has a long way to go” (p. 237) and that

“more fundamental research is needed to gather more knowledge on pronunciation variation.” Similarly, phonological variation in ordinary communicative speech is attracting the attention of psycholinguists, who have begun to explore the effects of variation on human word recognition and language production (Deelman and Connine, 2001; Donselaar et al., 1999; Gaskell and Marslen-Wilson, 1996, 1998; Utman et al., 2000; Weber, 2001).

The Buckeye corpus was created with researchers at both ends of the speech chain in mind. It is a resource with which to study variation and to assess its consequences for speech processing. For those interested in the phenomenon of variation itself, the speech files are phonemically labeled to make it easy to perform acoustic and phonological analyses. Such work should lead to a better understanding of variation. Our hope is that this information will assist speech recognition scientists in modeling variation and assist psycholinguists in studying how it affects perception.

Precisely because the corpus was created to serve a wide range of needs in the speech sciences community, it is somewhat unique. When it is released (see below), the Buckeye corpus will be the only publicly available corpus of spontaneous American English that has been phonemically labeled and accompanied by high-fidelity speech files. The two corpora that come closest are the TIMIT corpus (Fisher et al., 1987), which consists of 6300 words of read speech, and a subset of the Switchboard corpus (Greenberg, 1997), which has 35,000 words of phonemically labeled band-limited telephone speech.

3. Corpus creation

Forty talkers were from the Columbus, Ohio community. All were natives of Central Ohio (i.e., born in or near Columbus, or moved there no later than age 10). The sample was stratified for age (under 30 and over 40) and sex, and the sampling frame was limited to middle-class Caucasians. Past work suggests that such a sample is large enough to ensure that the interspeaker varia-

tion observed in the corpus is representative of the speech community (Fasold, 1990).

Talkers were recruited through advertisements in local newspapers and through referrals from other talkers between October 1999 and May 2000. They were invited to come to Ohio State University's main campus to have a conversation about everyday topics such as politics, sports, traffic, schools. After the interview, talkers were debriefed on the conversation's true purpose and all consented to having their speech used in research.

After a significant amount of piloting different protocols for eliciting large amounts of unmonitored speech, a modified sociolinguistic interview format was chosen. Interviews were conducted in a small seminar room by one of two interviewers, a 32-year-old male or a 25-year-old female. Talkers sat in a chair facing the interviewer and wore a head-mounted microphone (Crown CM-311A), which allowed considerable freedom of movement. The microphone was fed to a DAT recorder (Tascam DA-30 MKII, 48 kHz sampling rate) via a Yamaha MV 802 amplifier, where the signal level was monitored by the interviewer.

Upon arrival, talkers were told that the purpose of the study was to learn how people express "everyday" opinions in conversation, and that the actual topic was not important. Each interview began with a few questions about the talker concerning his/her age, place of birth, family make-up, etc. This information was found by the interviewers to lead to questions that easily elicited opinions, such as how Columbus has changed over the years, how families get along, how children should be raised, etc. These topics in turn offered opportunities for talkers to express other opinions. In order to elicit more conversation, the interviewer often challenged the talker with other points of view or asked for illustrations of alternative opinions. As the session proceeded, talkers became less inhibited and the interview approximated a friendly conversation usually within 5 or 10 min of its beginning. Interviews lasted from 30 to 60 min, with the latter being the target length. To control for the possible influences of the interviewer's sex, each interviewer met with half of the talkers in each sex/age group.

4. Corpus transcription and labeling

The recorded conversations were first orthographically transcribed into written English text. In addition to being used in phonemic labeling, the written version of the corpus enabled us to determine early on some of the characteristics of the corpus. A few are described here.

Talkers spoke a total of 306,652 words (i.e., tokens), and 9600 different words in all. Slightly more than half of the word tokens (57%) were function words, with the remainder being content words. The left two bars in Fig. 1 show how the words vary in number of syllables when combined across all talkers. As is true in other transcribed corpora of informal speech (Carterette and Jones, 1974; Svartnick and Quirk, 1980), 1-syllable words dominate the token count whereas 1-, 2-, and 3-syllable words constitute the majority of word types. The two right bars show these same data averaged over talkers to obtain a profile of central tendency. The token percentages for the "average talker" are virtually identical to those for all talkers combined (compare first and third bars), but the distribution of word types changes across them: the percentage of 3-syllable words increased from the average talker to all talkers, whereas the percentage of 1-syllable words did the reverse. This suggests that as a group, talkers spoke more unique 3-syllable than 1-syllable words, so that when the 3-syllable words were combined over talkers, they constituted a larger percentage of word types. It is worth mentioning that the averaged talker data, when measured in percentages as in Fig. 1, are very representative of all 40 talkers, as the standard deviation was less than 2% in all five word-length categories. Such low variability is surprising given how much talkers differed from one another in how much they said (range: 3100–12,200 words).

To phonemically label the corpus, it was necessary to develop a protocol to standardize labeling. We began by adopting the TIMIT labeling guidelines and using the DARPA phonetic alphabet (Garofolo et al., 1993; Seneff and Zue, 1988). Members of the research team independently labeled 1 min stretches of speech (~200 words) by a talker. Choice of phonemic labels and their boundaries were then compared and differences discussed. This

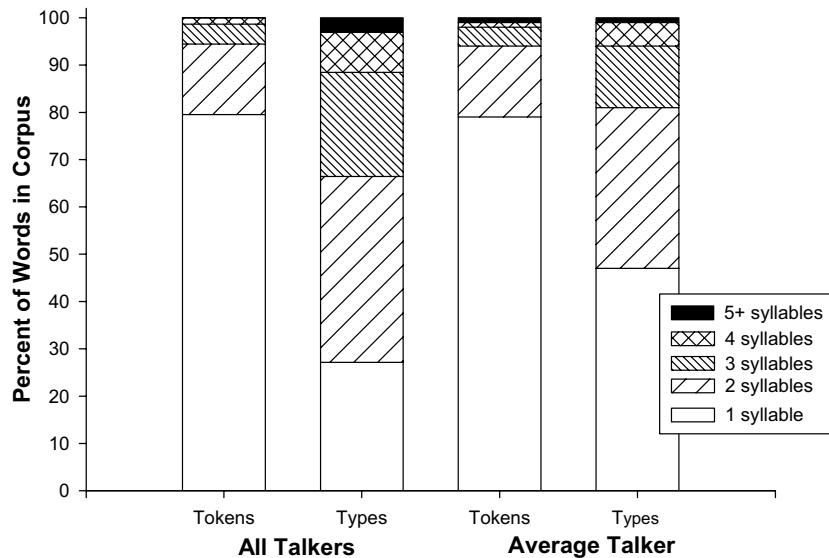


Fig. 1. Token and type characteristics in the Buckeye corpus.

process was repeated many times on new stretches of speech and across multiple talkers. A labeling protocol and phonemic alphabet emerged from these efforts. Both are contained in a manual that was written for transcribers, a copy of which can be found at <http://vic.psy.ohio-state.edu>.

Two passes are made through each speech file during labeling. In the first, labeling is automated. A small (1 min) stretch of speech (downsampled to 16 kHz) and its corresponding orthographic transcription are fed into Entropics Aligner, which assigns phonemic labels using acoustic phone models that were built from training on the TIMIT corpus of spoken English. The second pass through a speech file involves hand-correcting the phonemic labels and their alignment. Once completed, phonemic labeling of what was said along with the acoustic properties of each phone can be analyzed. To date, first-pass alignment has been completed on 64% of the corpus (193,000 words) and the more time-consuming second-pass alignment has been completed on 35% (104,000 words).

5. Test of labeling consistency

A test of labeling consistency was conducted to measure inter-transcriber agreement in phone

labeling. Four 1 min samples (220 words each) of speech were independently labeled by four transcribers. The samples were taken from the middle of the interviews of four talkers, one from each age/sex category. The talkers had not previously been labeled by any of the transcribers, and each transcriber worked alone on the samples rather than using “consensus-based” labeling (Shriberg and Lof, 1991).

Agreement was measured by counting the number of phone labeling agreements for all pairs of transcribers (six total). For example, if a particular phone was labeled by the first transcriber (T1) as “ih” [I], by the second transcriber (T2) as “ix” [i], and by transcribers three and four (T3 and T4) as “ih” [I], then the number of transcriber pairs who agree with each other is three (T1–T3, T1–T4, T3–T4) and the number of transcriber pairs who disagree with each other is also three (T1–T2, T3–T2, and T4–T2). With one of the transcribers labeling the phone differently from the others, only three out of the six possible pairings of transcriptions show agreement, making the agreement rate 50% (i.e., agreement = agree/(disagree + agree)). This method of calculating agreement is more conservative than that used by Shriberg and Lof (1991), who penalized disagreements less by dividing the number of disagree-

ments by two in the preceding formula, which gives an agreement rate of 66% in this example.

There were 2159 phones in the passages labeled for the test. With six transcriber pairs per phone, this resulted in 12,954 paired comparisons of the labels. Of these, 10,402 pairs (80.3%) showed agreement. The second column in Table 1 shows the percentage of transcriber agreements summed over all segments and then broken down by broad segment classes. Listed in column three are values of the *kappa* statistic, a widely used measure of interjudge agreement that controls for chance and varies between 0 and 1.0 (Cohen, 1960; Cucchiaroni, 1996; Perreault and Leigh, 1989). Larger values indicate greater agreement. A related measure, *max(kappa)*, is listed in the next column and reflects the consistency of phonetic symbol use across transcribers, with larger values indicating greater consistency.

Transcribers agreed most often when labeling stops and fricatives. Agreement dropped slightly when labeling nasals and liquids. The high *max(kappa)* values across these four consonant categories indicates that transcribers were using the symbol set similarly.

Less labeling agreement was found with the vowels. The cause of this drop in labeling consistency can be better understood by examining the labeling matrix for monophthongal vowels (Table 2; diphthongs were labeled with a high degree of accuracy). The most reliably transcribed monophthongs included the point vowels “uw” [u] 78% agreement, “ow” [o] 69% agreement, “iy” [i] 67%

agreement, “aa” [a] 64% agreement, and “ae” [æ] 63% agreement. Two vowel symbols that showed extremely low levels of transcriber agreement were “ix” [ɨ] 17% and “ux” [ʌ] 7%. Such low agreement, together with a drop in the *max(kappa)* value for vowels, indicates that the transcribers used these symbols inconsistently. Comparison of labels showed that one transcriber used “ux” fairly often while the others used this symbol only rarely, choosing instead either “ix” or “ax” for the same segments. To examine the effects of this variation in symbol usage on labeling agreement, we recalculated the measures in Table 1, collapsing “ix” [ɨ] and “ux” [ʌ] onto the more frequently used symbols for such segments (“ih” [ɪ] and “ax” [a] respectively). The revised estimates of transcriber reliability are shown in the last two rows of the table. All measures of transcriber reliability improved noticeably from this slight broadening of the labeling symbol set. It would continue to improve if other confusable vowels were combined as well (e.g., “ah” [ʌ] and “ax” [a]).

Despite the lower level of labeling agreement found with the vowels, the results of this test of labeling consistency compare favorably with other studies, even though most of these used read speech rather than spontaneous speech, a few of which were in German (Amorosa et al., 1985; Burkowski, 1967; Eisen, 1991; Irwin, 1970; Philips and Bzoch, 1969). For example, using a similarly broad symbol set in transcribing read speech, Eisen (1991) found labeling accuracy was 88% for obstruents, 93% for sonorants, and 83% for vowels. The corresponding values in the current test were 92%, 87%, and 74%, respectively.

In addition to measuring reliability of phonemic labeling, we also compared transcribers’ temporal placement of labels. Transcribers could hypothetically choose exactly the same phonemic labels yet mark the boundaries of segments at different locations in time in the speech waveform. The average deviation of the boundary locations were calculated for those segments for which transcribers chose the same phonetic symbol (i.e., unanimous agreement; see Table 1). The mean deviation in boundary placement across all six pairs of transcribers was 16ms. This degree of variation is comparable to that found by others. For example,

Table 1
Measures of labeling consistency in the Buckeye corpus

	N	% Agree	Kappa	Max (kappa)	% Unanimous
Overall	2159	80.3	.797	.926	62
Stops	368	92.9	.918	.965	74
Fricatives	507	91.2	.894	.947	76
Nasals	331	87.5	.82	.942	68.5
Liquids	251	86.5	.802	.927	56
Vowels	907	69	.66	.87	49
Vowels (~ix,ux)	907	73.6	.701	.89	53
Overall (~ix,ux)	2159	82.2	.816	.933	63.5

Table 2
Vowel transcription matrix for monophthongal vowels

	iy[i]	ih[I]	eh[ɛ]	ae[æ]	ix[ɪ]	ax[a]	ah[ʌ]	ux[ʊ]	uw[u]	uh[U]	ow[o]	ao[a]	aa[a]
iy	468												
ih	180	478											
eh	0	65	337										
ae	0	8	69	222									
ix	41	192	38	0	84								
ax	2	67	64	17	115	274							
ah	0	9	20	9	4	111	214						
ux	3	19	8	0	25	34	8	8					
uw	0	9	0	0	2	0	0	6	81				
uh	1	6	0	0	5	1	2	6	6	37			
ow	0	0	9	0	2	15	24	4	0	0	198		
ao	0	0	0	0	0	1	6	0	0	3	34	96	
aa	0	0	5	26	0	20	43	0	0	0	2	37	232
%	67	46	55	63	17	38	48	7	57	55	69	67	64

Agreements are along the diagonal and disagreements on the off diagonal. The bottom row contains the percent agreement for each vowel.

Wesenick and Kipp (1996) found in broad phonemic labeling of read sentences by three transcribers that the average deviation in label placement was about 10ms. Interestingly, this value almost doubled (to 18ms) when hand labeling was compared with machine labeling.

The labeling consistency results are presented not only for the reader to evaluate the success of our labeling methodology, but also to underscore the realities of studying spontaneous speech using phonemic labels. Speech varies tremendously and even unpredictably when spoken in a relaxed style. Labeling consistency will reflect this. Researchers interested in studying the labeled corpus should do so in the context of these results. Indeed, researchers interested in the relationship between the acoustic properties of seemingly impoverished speech and its perception will likely find the variability in labeling of interest.

6. Corpus availability

Once 50% of the corpus has been transcribed, it will be made available free of charge to the research community through the Linguistics Department at Ohio State University, which at present we anticipate will be in the Fall of 2005. The release will include the speech files, accompanying ortho-

graphic and phonemic transcriptions, software to search the corpus, the transcription manual, and data on transcriber consistency tests. The entire corpus will be released when labeling is finished.

7. Conclusions

Although the Buckeye corpus was created to study phonological variation, it should be useful to scientists interested in many other aspects of spontaneous speech. Indeed, a guiding principle in its creation was to make it as versatile as possible. This is one reason why the speech of talkers in different age/sex strata were sampled and why the entire corpus is being phonemically labeled. In the future, we hope to enhance the corpus even further by including additional information with the speech files, such as prosody (the speech of one talker has already been ToBI transcribed) and talking style.

Acknowledgments

We thank Robin Dautricourt, Craig Hilts, Matthew Makashay, Jeff Mielke and the many undergraduates for assistance in transcription. This work was supported by NIDCD Grant 004330.

References

- Amorosa, H., von Benda, U., Wagner, E., Keck, A., 1985. Transcribing phonetic detail in the speech of unintelligible children: a comparison of procedures. *Brit. J. Disorders Comm.* 20, 281–287.
- Brown, G., 1990. *Listening to Spoken English*. Longman, New York, NY.
- Burkowsky, M.R., 1967. A study of the perception of adjacent fricative consonants. *Phonetica* 17, 38–45.
- Byrd, D., 1992. Preliminary results on speaker-dependent variation in the TIMIT database. *J. Acoust. Soc. Amer.* 92, 593–596.
- Byrd, D., 1993. 54,000 American stops. *UCLA Work. Papers Phonet.* 83, 97–116.
- Byrd, D., 1994. Relations of sex and dialect to reduction. *Speech Comm.* 15, 39–54.
- Carterette, E.C., Jones, M.H., 1974. *Informal Speech: Alphabetic and Phonemic Texts with Statistical Analyses and Tables*. University of California Press, Berkeley, CA.
- Cohen, J., 1960. A coefficient of agreement for nominal scales. *Educ. Psychol. Meas.* 20, 37–46.
- Cucchiari, C., 1996. Assessing transcription agreement: methodological aspects. *Clin. Ling. Phonet.* 10, 131–155.
- Dalby, J.M. 1986. *Phonetic structure of fast speech in American English*. Unpublished doctoral dissertation, Indiana University.
- Deelman, T., Connine, C.M., 2001. Missing information in spoken word recognition: nonreleased stop consonants. *J. Exp. Psychol.: Human Percept. Perform.* 27, 656–663.
- Donselaar, W., Kuipers, C., Cutler, A., 1999. Facilitory effects of vowel epenthesis on word processing in Dutch. *J. Mem. Lang.* 41, 59–77.
- Eisen, B., 1991. Reliability of speech segmentation and labelling at different levels of transcription. *Eurospeech 1991*, 673–676.
- Fasold, R.W., 1990. *The Sociolinguistics of Language*. Blackwell Publishers, Oxford.
- Fisher, W., Zue, V., Bernstein, J., Pallet, D., 1987. An acoustic-phonetic data base. *J. Acoust. Soc. Amer.* 81, S92–S93.
- Garofolo, J., Lamel, L., Fisher, W., Fiscus, J., Pallet, D., Dahlgren, N., 1993. *Darpa, TIMIT, Acoustic-phonetic continuous speech corpus*. (NISTIR Publication No. 4930). Washington, DC: US Department of Commerce.
- Gaskell, G., Marslen-Wilson, W.D., 1996. Phonological variation and inference in lexical access. *J. Exp. Psychol.: Human Percept. Perform.* 22, 144–158.
- Gaskell, G., Marslen-Wilson, W.D., 1998. Mechanisms of phonological inference in speech perception. *J. Exp. Psychol.: Human Percept. Perform.* 24, 380–396.
- Gimson, A.C., 1989. *An Introduction to the Pronunciation of English*. Edward Arnold, New York.
- Greenberg, S., 1997. *The Switchboard transcription project*. In *Research Report #24, Large Vocabulary Continuous Speech Recognition Summer Research Workshop Technical Report Series*. Center for Language and Speech Processing, Johns Hopkins University, Baltimore, MD.
- Guy, G.R., 1980. Variation in the group and the individual: the case of final stop deletion. In: Labov, W. (Ed.), *Locating Language in Time and Space*. Academic Press, New York.
- Irwin, R.B., 1970. Consistency of judgments of articulatory productions. *J. Speech Hear. Res.* 13, 548–555.
- Jurafsky, D., Bell, A., Fosler-Lussier, E., Girand, C., Raymond, W., 1998. Reduction of English function words in Switchboard. In: *Proceedings of the International Conference on Spoken Language Processing (ICSLP-98)*.
- Jurafsky, D., Bell, A., Gregory, M.L., Raymond, W.D., 2001. Probabilistic relations between words: evidence from reduction in lexical production. In: Bybee, J., Paul, H. (Eds.), *Frequency and the Emergence of Linguistic Structure*, John Benjamins, Amsterdam, pp. 229–254.
- Keating, P., Byrd, D., Flemming, E., Todaka, Y., 1994. Phonetic analyses of word and segment variation using the TIMIT corpus of American English. *Speech Comm.* 14, 131–142.
- Labov, W., 1994. *Principles of Linguistic Change*. Blackwell, Cambridge, MA.
- Manuel, S.Y., Shattuck-Hufnagel, S., Huffman, M., Stevens, K.N., Carlson, R., Hunnicut, S., 1992. Studies of vowel and consonant reduction. In: Ohala, J.J., Nearey, T.M., Derwing, B.L., Hodge, M.M., Wiebe, G.E. (Eds.), *Proceedings of the 1992 International Conference on Spoken Language Processing*, pp. 943–946.
- Neu, H., 1980. Ranking of constraints on /t,d/ deletion in American English: a statistical analysis. In: Labov, W. (Ed.), *Locating Language in Time and Space*. Academic Press, New York.
- Perreault, W.D., Leigh, L.E., 1989. Reliability of nominal data based on qualitative judgements. *J. Market. Res.* 26, 135–148.
- Philips, B.J.W., Bzoch, K.R., 1969. Reliability of judgements of articulation of cleft palate speakers. *Cleft Palate J.* 6, 24–34.
- Seneff, Zue, V., 1988. In the TIMIT CDROM documentation. *Linguistic Data Consortium, Philadelphia*.
- Shockey, L., 1973. *Phonetic and phonological properties of connected speech*. Doctoral dissertation, Ohio State University.
- Shriberg, L.D., Lof, G.L., 1991. Reliability studies in broad and narrow phonetic transcription. *Clin. Ling. Phonet.* 5, 225–279.
- Strik, H., Cucchiari, C., 1999. Modeling pronunciation variation for ASR: a survey of the literature. *Speech Comm.* 29, 225–246.
- Svartvick, J., Quirk, R., 1980. *A Corpus of English Conversation*. Gleerup, Lund.
- Utman, J.A., Blumstein, S.E., Burton, M.W., 2000. Effects of subphonemic and syllable structure variation on word recognition. *Percept. Psychophys.* 62, 1297–1311.
- Weber, A., 2001. Help or hindrance: how violation of different assimilation rules affects spoken-language processing. *Lang. Speech* 44, 95–118.
- Wesenick, M.-B., Kipp, A., 1996. Estimating the quality of phonetic transcriptions and segmentations of speech signals. In: *Proceedings of the ICSLP, Philadelphia, USA*, pp. 129–132.