An automatic close copy speech synthesis tool for large-scale speech corpus evaluation

Dafydd Gibbon[1], Jolanta Bachan[2]
[1]Universität Bielefeld
Postfach 100131, D-33501 Bielefeld, Germany
[2]Uniwersytet im. Adama Mickiewicza
H. Wieniawskiego 1, 61-712 Poznań, Poland
E-mail: gibbon@uni-bielefeld.de, jolabachan@gmail.com

Abstract

The production of rich multilingual speech corpus resources on a large scale is a requirement for many linguistic, phonetic and technological tasks, in both research and application domains. It is also time-consuming and therefore expensive. In particular the human component in the resource creation process is prone to inconsistencies, a situation which has frequently been documented in studies of cross-transcriber consistency in manual time-aligned signal annotation. In the present case, corpora of three languages were to be evaluated and corrected: (1) Polish, a large automatically annotated and manually corrected single-speaker TTS unit-selection corpus in the BOSS Label File (BLF) format, (2) German and (3) English, the second and third being manually annotated multi-speaker story-telling learner corpora in Praat TextGrid format. A method is provided for supporting the evaluation and correction of time-aligned annotations for the three corpora by permitting a rapid audio screening of the annotations by an expert listener for the detection of perceptually conspicuous systematic or isolated errors in the annotations. The criterion for perceptual conspicuousness was provided by converting the annotation formats into the interface format required by the MBROLA speech synthesiser. The audio screening procedure is complementary to other methods of corpus evaluation and does not replace them. Conceptually the ACCS synthesis tool is intended as an extension of the BLARK toolkit for speech corpora.

1. Efficient quality control of richly annotated corpora

The production of rich multilingual speech corpus resources on a large scale is a requirement for many linguistic, phonetic and technological tasks, in both research and application domains. It is also time-consuming and therefore expensive. In particular the human component in the resource creation process is prone to inconsistencies, a situation which has frequently been documented in studies of cross-transcriber consistency in manual time-aligned signal annotation, particularly with prosody (Grice 2006; Gibbon et al. 1997; Gibbon et al. 2000). In the present case, corpora of three languages were to be evaluated and corrected: (1) Polish, a large automatically annotated and manually corrected single-speaker TTS unit-selection corpus in the BOSS Label File (BLF) format (Demenko et al. 2006), (2) German and (3) English, the second and third being manually annotated multi-speaker story-telling learner corpora in Praat TextGrid format (Boersma 2001; Gut et al. 2004).

The first general goal is to provide a method for supporting the evaluation and correction of time-aligned annotations for the three corpora by permitting rapid audio screening of the annotations by an expert listener, who detects perceptually conspicuous systematic or isolated errors in the annotations. The criterion for perceptual conspicuousness is provided by converting the annotation formats into the interface format required by a suitable speech synthesizer, in this case the PHO format required by the MBROLA synthesizer (Dutoit & al. 1996). The audio screening procedure is complementary to other methods of corpus evaluation and does not replace them. Functionally, the ACCS synthesis tool is intended as an addition to the BLARK (Krauwer 2005) toolkit for speech corpora.

A second general goal is a practical one: the method is also intended for use with corpora for less-resourced languages and for use in areas with very basic infrastructures. The method should therefore not only be of good quality and well-defined, but at the same time straightforward and as far as possible not dependent on complex cutting edge tools, expensive software packages, or the internet. Otherwise, usability by development teams working with sub-optimal infrastructures in an under-resourced languages paradigm is not assured.

The focus in the present application is on segmental annotation evaluation, but the pitch pattern of the original speech signals was also extracted and mapped into the synthesiser interface in order to provide as natural a re-synthesis as possible for the segmental evaluation. Prosodic annotation is not evaluated. The issues addressed are:

1. Consistency of labels used with defined label set (e.g. phoneme or phone set).
2. Correct time-stamp assignment (e.g. segment duration).
3. Correct label selection from the relevant inventory.

The first of these problems is a 'syntactic' issue which can be dealt with automatically, given a specified inventory of labels. The second and third are 'semantic' issues which require an element of subjective assessment, in that mapping of the annotation to the speech signals is involved. This assessment can be cross-transcriber checking; the present approach uses re-synthesis of the annotated corpora, with preliminary quality checking based on standard perceptual tests. Two modes of operation are used:

1. Continuous listening for comprehensibility and naturalness.
2. Comparative listening to the original and the re-synthesis, for accuracy.

2. Requirements

Starting with the general specifications already outlined, a procedure of subjective diagnostic annotation evaluation was defined, incorporating a tool which applies Automatic Close Copy Speech (ACCS) synthesis to a time-aligned annotated corpus and whose output is then pre-checked by means of standard perceptual testing procedures (Gibbon et al. 1997), before being used by the expert listener for direct subjective audio screening of annotation quality. The ACCS technique itself is well-known (Bachan & Gibbon 2006; Bachan 2007), but extensive, consistent and well-defined application to the corpus annotation evaluation task does not figure in the literature, and no generic tool for this purpose was previously available.

The following specific requirement specifications were formulated:

1. The system should provide a multi-platform language-independent ACCS synthesis shell with phone inventories and annotation formats to be defined for each application case.
2. The system and its components should be freely available for academic purposes.
3. A readily available and easy-to-use speech synthesis system should be used.
4. The speech synthesis process should be fully automatised (bar the two case-specific definitions mentioned under #1).
5. The ACCS synthesiser should be evaluated initially with perception tests, then applied by an expert listener to the evaluation of the annotated corpora.

3. Design

A conventional Text to Speech (TTS) synthesis architecture has two main components: the Natural Language Processing Component (NLP) and the Digital Signal Processing Component (DSP). In ACCS synthesis, information from the annotated speech corpus takes the place of the entire NLP front-end of the TTS system. The main tasks of the NLP front-end are replaced fairly straightforwardly as follows (Bachan & Gibbon 2006):

1. *Phonetisation model:* replaced by a phoneme inventory based validation module designed for phonemically annotated corpora, as in the present case; the module presupposes forced alignment pre-processing to provide phoneme-level annotation in the case of orthographic, syllable-sized etc. annotations.
2. *Duration model:* from the time-stamps of the annotation (details dependent on annotation format).
3. *Pitch model:* pitch extraction algorithm over the given label time domains.

The modules are cascaded in the order *Phonetisation – Duration – Pitch*. The input is a pair of a speech signal file and a time-aligned phonemic annotation, followed by phoneme validation, followed by duration extraction, followed by pitch extraction, and finally by integration of the phoneme labels, durations and pitch positions and values into the synthesizer interface format (MBROLA PHO format). The main data flow steps are shown in Figure 1.

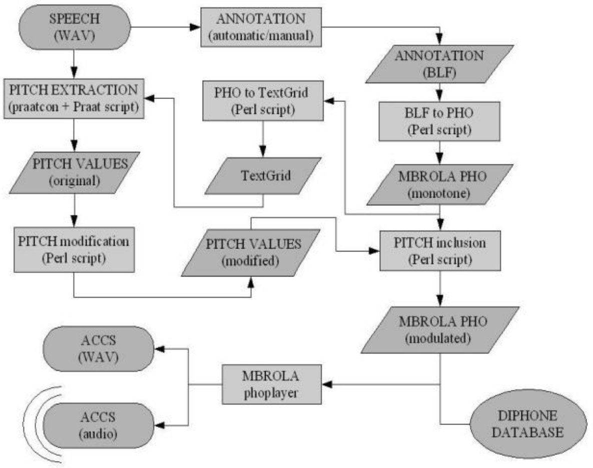

Figure 1: ACCS re-synthesis implementation data flow.

4. Implementation

The TTS signal processing engine selected for the purpose is MBROLA (Dutoit et al. 1996), a *de facto* standard for diphone synthesis, rather than the more complex and (in principle) higher quality variable unit selection type of engine. The design decisions behind the selection of MBROLA, despite its age, limitation to diphone synthesis, and restricted dynamic range, are:

1. Conformance to the requirement for free availability for academic purposes (engine and suitable voices, i.e. diphone databases).
2. Clear and simple interface for inserting label, duration and pitch information.
3. Simple interfacing to standard Perl and UNIX scripting techniques for rapid prototyping.

With one exception, the modules are implemented in Perl; the exception is the pitch extraction module, which is implemented as a Praat script. The reason for selecting a scripting environment, aside from its value for rapid prototyping, is the transparency of scripting language techniques for less experienced developers of resources for technologically less resourced languages.

The reference formats for interfacing within the system are the Praat TextGrid format (a flat tier hierarchy of attribute-value pairs), and the MBROLA PHO format (a list of tuples of phoneme label, duration in milliseconds, and an optional series of pairs of pitch position in percent of the segment duration and F0 value in Hz).

The modules of the ACCS system are listed in Table 1.

The following resources are required for the deployment of the ACCS system:

1. Perl software (duration processing; formatting).
2. Praat software (pitch extraction).
3. MBROLA runtime engine. (speech output).

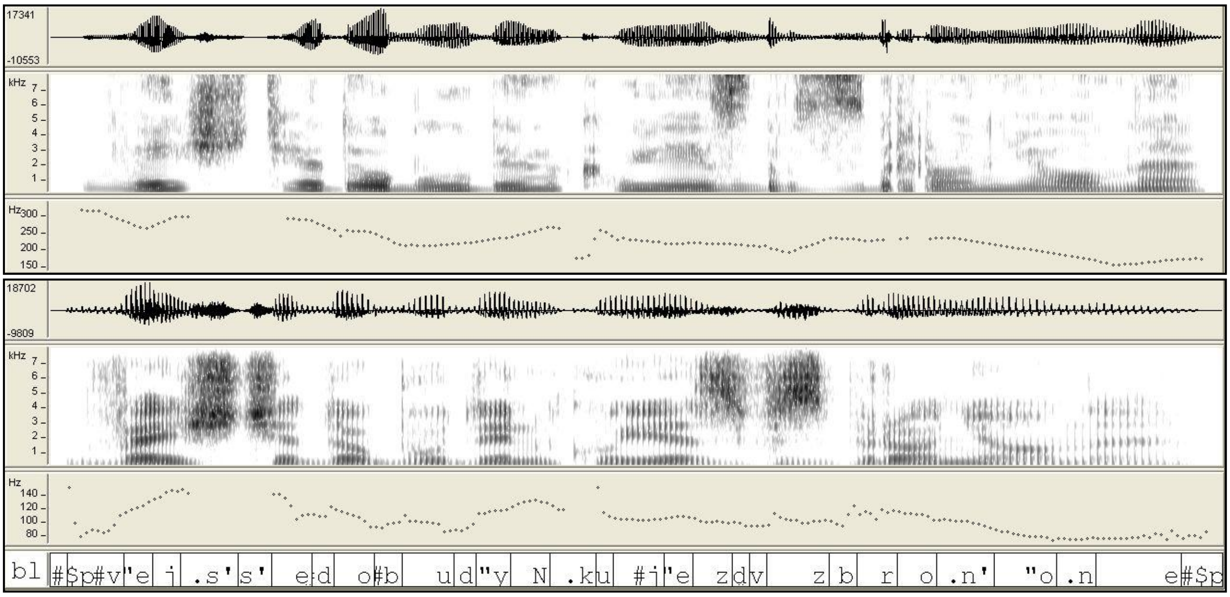

Figure 2: ACCS synthesis (top) vs. original recording file with annotation (bottom) for the Polish utterance "*Wejście do budynku jest wzbronione*" ("No admittance to the building").

4. MBROLA diphone database (voice) for the language concerned.
5. Corpus:
   a. Speech signal (in WAV format).
   b. Time-aligned phonemic annotation (are converted to Praat TextGrid format).

The ACCS system runs in Linux, Windows and Mac environments, and will be made available under open source conditions. Voice-making software is required for applications to other languages. An example from the corpus, with the ACCS synthesised file (top) and the original recording (bottom) are shown for comparision in Figure 2.

User interface:	Schedule selection.
Schedule control:	Handling of multiple annotations and multiple voices.
Data processing:	Perl: Annotation format to MBROLA interface format conversion. Praat: F0 pattern extraction for each phone interval. Perl: F0 pattern insertion into MBROLA interface file.
Output:	MBROLA: Mapping of interface file to WAV file, and audio output.

Table 1: Main modules of ACCS system.

5. Evaluation

Preliminary benchmarking and evaluation of the ACCS procedure was performed on the manually checked automatically annotated Polish data (Bachan 2007) by means of a set of short perception tests based on EAGLES standards (Gibbon et al. 1997). The three speech quality tests were administered to 19 Polish subjects and 2 non-native speakers with a good command of Polish. The youngest Polish subject was 8 years old, the oldest 55. The non-natives' ages were 22 and 25. The subjects took the tests separately. The stimuli were played to the subjects once, or twice if the subject requested this. Each test session lasted 30-45 min. The tests investigated:

1. *Test 1:* sentence and word recognition in 10 semantically predictable and 10 semantically unpredictable sentences, with repetitions.
2. *Test 2:* mean opinion score (MOS) of subjective sentence quality on a five-point rating scale.
3. *Test 3:* intonation of isolated words.

5.1 Test 1: sentence and word recognition

Method: The synthetic material is presented to the subjects. Their task is to write down what they hear in the answer sheet. A set of semantically unpredictable sentences is used in order to eliminate the influence of the top-down processing on the perception task (Clark & Yallop 1995: 312, Ryalls 1996: 94).

Material: 10 meaningful and 10 meaningless sentences synthesised with the Polish female MBROLA voice (PL1) (Szklanny & Marasek 2002).

Results: The results of Test 1 on sentence and word recognition are presented in Table 2. Despite the small test sets, the results show that ACCS is highly intelligible by the Polish native speakers, though not by the non-natives. The non-natives' poor results are apparently due to the fact that although the non-natives spoke vernacular Polish well, they were unfamiliar with the vocabulary used in the tests. Therefore, after testing with two foreign subjects, no more test with non-native speakers were carried out.

In Table 3, a comparison of the intelligibility of sentences and words in semantically predictable and unpredictable sentences is shown. The results indicate, predictably, that semantic information helps in speech signal recognition; both sets of sentences get high scores on word perception under this condition.

	N	*Sentences (absolute)*	*Sentences (percent)*	*Words (absolute)*	*Words (percent)*
Polish male	8	13.63	68	111.00	88
Polish female	11	14.36	72	115.82	92
Polish overall	19	14.05	70	113.79	90
Non-natives	2	2.00	10	61.00	48

Table 2: Results for Test 1 – average correctly recognised units in 20 sentences (126 words). *N*: number of subjects.

	N	*Predictable*	*N*	*Unpredictable*
Sentences	10	83.30%	10	55.30%
Words	75	96.28%	51	81.53%

Table 3: Comparison of results for semantically predictable and unpredictable sentences, Polish subjects only. N is the number of items.

5.2 Test 2: subjective sentence quality

Method: The subjects are asked to evaluate the quality of isolated long (multiple) sentences from the corpus on a five-point MOS scale: *Excellent – Good – Fair – Poor – Bad*.

Material: 10 different compound sentences have been chosen from the Polish corpus, which has a male voice. 15 compound sentences are synthesised using the Polish female MBROLA voice (PL1) (Szklanny & Marasek 2002). In order to handle the mismatch between the corpus and voice genders, the following procedure was used:

1. In 10 sentences, the pitch value was adapted straightforwardly for a female voice by raising by one octave. This voice is called *pseudo-female*. This procedure is not a full voice-morphing, but only modification of pitch.
1. 5 sentences from the set have the original pitch values extracted from the recordings of the male speaker. This voice is called *pseudo-male*.

Additionally, 5 non-synthesised sentences from the original corpus, uttered by the male speaker, were included.

Altogether, 20 sentences were used in the test, but only 10 different sentences. The set is kept small so as not to tax the volunteer listeners unduly, but were selected to be representative for the task. The sentences were played in a random order.

Results: The results of Test 2 on subjective sentence quality are presented in Table 4. In the test the original voice received the best scores, as expected. Both Poles and non-native speakers graded it highly. The range of the grades given to the original voice is the same in both groups with the minimal grade 3 and the maximal grade 5. The adapted synthetic pseudo-female voice received much worse scores from both groups of subjects, again. It was graded approximately two points less (in the five-point rating scale) than the original voice. This result did not vary significantly from the score which the synthetic pseudo-male voice received in the evaluation by Poles. The pseudo-male voice scored just almost half a point less (0.44) than the pseudo-female voice. However, non-natives graded the pseudo-male voice worse by one point.

The results show that the subjects graded the human voice much better than the synthetic voices. But it has to be underlined that the synthetic voice was confronted with recordings of a professional speaker recorded in a professional studio. As expected, the pseudo-female voice was evaluated better than the pseudo-male voice, although the scores were not very different. This suggests that the pseudo-male voice seemed approximately as natural and intelligible as the pseudo-female voice.

The considerably worse results of the pseudo-female and pseudo-male voices in comparison with the original voice may suggest that the articulation in the synthetic signal was not very good. Whether the problem might lie in the annotations or in the diphone database itself was not investigated further at this stage.

After the testing procedure the subjects were asked informal questions about what they heard and it turned out that they did not realise they were exposed to synthetic speech. They believed that they were evaluating the articulation of human speakers. These informal observations are very encouraging in respect of the naturalness of the re-synthesis procedure.

	N	*Original*			*Pseudo-female*			*Pseudo-male*		
		MOS score/5	*STDV*	*Max: Min*	*MOS score/5*	*STDV*	*Max: Min*	*MOS score/5*	*STDV*	*Max: Min*
Polish male	8	4.70	0.52	5:3	3.03	0.83	4:1	2.45	0.75	4:1
Polish female	11	4.69	0.50	5:3	2.50	0.87	4:1	2.18	0.89	4:1
Polish overall	19	4.69	0.51	5:3	2.73	0.89	4:1	2.29	0.85	4:1
Non-native	2	4.50	0.71	5:3	2.30	0.81	4:1	1.30	0.48	2:1

Table 4: Test results for Test 2. N is the number of subjects, MOS score/5 stands for Mean Opinion Score out of 5, STDV is the standard deviation, Max:Min are the maximal and minimal scores given by the population.

5.3 Test 3: isolated word intonation

Method: Isolated synthetic words are presented to the subjects. The subject assesses whether the words would appear at the end of a statement or at the end of a question based on the intonation of the word.

Material: A set of 20 words cut out of the whole sentences and synthesised using the Polish female MBROLA voice (PL1) (Szklanny & Marasek 2002). These words originally appeared at the end of a statement, a question, an exclamation and at the end of a continuation phrase with rising intonation. The words from the exclamations and continuation phrases are distractors, and are not the subject of the study. But it is assumed that exclamation words will be recognised as statement words because of their falling intonation, and continuation phrase-words will be recognised as question words because of their rising intonation.

The words are presented in random order.

		INPUT			
		Statement	*Question*	*Exclamation*	*Continuation phrase*
JUDGEMENTS	**Polish male** *(N=8)*				
	St	80%	5%	90%	68%
	Q	0%	85%	8%	10%
	DK	20%	10%	3%	23%
	Polish female *(N=10)*				
	St	88%	6%	70%	68%
	Q	6%	86%	17%	16%
	DK	6%	8%	16%	16%
	Polish overall *(N=18)*				
	St	84%	6%	79%	68%
	Q	3%	86%	11%	13%
	DK	12%	9%	10%	19%
	Non-natives *(N=2)*				
	St	40%	50%	20%	30%
	Q	20%	50%	70%	30%
	DK	40%	0%	10%	40%

Table 5: Test results for Test 3. Judgements given by the subjects are in rows, actual input is in columns. There were 5 items for each input category. St – Statements, Q – Question, DK – Don't know. N is the number of subjects.

Results: The test results for Test 3 are presented in Table 5. The results for the Polish listeners show that:

1. 84% of the statement words were recognised correctly as words at the end of a statement,
2. 86% of the question words were recognised correctly as words at the end of a question,
3. 79% of the exclamation words were recognised as words at the end of a statement, indicating that the intonation of these words was similar to the intonation of a statement, as expected.
4. 68% of the continuation phrase words were recognised as words at the end of a statement. This result was not expected, because it was assumed that the intonation pattern of words at the end of continuation phrases will sound more like a question than a statement. However, the subjects had the biggest problem assessing the category of the continuation phrase words and 19% of them were marked as "Don't know."
5. 3% of the statement words were recognised incorrectly as words at the end of a questions; 6% of question words were recognised incorrectly as words at the end of a statement.

To sum up, the overall results for Poles of the correctly recognised statement-words and question-words indicate that the intonation in the ACCS synthesis system is good. Exclamation words and continuation phrase words were added to the test as distractors and were not the main objective of this study.

Discussion. The main annotation error types remaining in the conventionally corrected Polish corpus were:

1. incorrect phoneme labels, both out-of-inventory ('syntactic') and wrong selection ('semantic');
2. incorrect final pause boundary ('semantic'; due to a bug in the automatic annotator).

Other types of time-stamp error were rare; the annotation had been carefully checked, but it is also possible that listener toleration of duration variability is high.

6. Audio screening

The preliminary speech output assessment tests carried out on the ACCS synthesis output not only showed that this kind of synthetic speech is of good quality, but also showed that the corrected annotations themselves are accurate, within the limits of perceptual tests. The automatically close copied speech used in the audio screening relies heavily on the annotations and their quality is reflected in the performance of the ACCS system and the synthetic speech output itself. This result demonstrated the potential of the ACCS system as a proof of concept for the evaluation of other corpora.

Although the Polish data had been automatically segmented and labelled, and carefully manually edited, numerous errors were still detected using the audio screening technique. However, in relation to the size of the corpus, the number of errors was reasonably low. The ACCS synthesis system was adapted for the English and German data; processing of these data is currently in progress, but since the German and English data were only manually annotated, the ongoing task of identifying and manually correcting transcriber errors is much greater.

7. Conclusion

The goals set in the requirement specification were met. The ACCS audio screening method provided both a proof of concept of the workability of the approach in providing a rapid overview of the quality of the time-aligned annotations, and proved its practical value in uncovering many additional errors which had been missed by standard methods of manual checking by trained personnel.

It is clear that the prototype would benefit from additional statistical testing. However, it is not clear what difference this would make to the actual audio screening application. On the ergonomic side, it became clear that the current prototype would benefit from enhancement with an interactive GUI-based control and error logging system in order to enable rapid changes to be made to the annotations and also from an interface with annotation software for efficient error correction.

In summary, the ACCS system has turned out to be a valuable tool for rapidly obtaining an overview of annotation quality, with respect to identifying both systematic errors and random slips. Other uses of the system are currently under development, including the evaluation of prosodic annotation, prosodic parameter manipulation for corpus-based speaker attitude modelling, and the training of labelling personnel.

8. Acknowledgements

We are grateful to Grażyna Demenko, Linguistics Department, U Poznań, and to Thorsten Trippel and Catharine Oertel for feedback on applications of the ACCS system in the LiteraSyn Project funded by Universität Bielefeld. The work was partly funded by an ERASMUS exchange grant to Jolanta Bachan, 2006-2007, partly by the LiteraSyn project, and partly by the Polish nationally funded project KBN 3T11C 038 27 Corpus based speech synthesis based on suprasegmental selection, 2004-2006, directed by Grażyna Demenko.

9. References

Bachan, J. & Gibbon, D. (2006). Close Copy Speech Synthesis for Speech Perception Testing. In: *Investigationes Linguisticae*, vol. 13, pp. 9--24.

Bachan, J. (2007). Close Copy Speech Synthesis for Perception Testing and Annotation Validation. M.A. thesis. Adam Mickiewicz University, Poznań, Poland.

Boersma, P. (2001). Praat, a system for doing phonetics by computer. *Glot International* 5:9/10, pp- 341--345.

Clark, J. & Yallop, C. (1995). *An Introduction to Phonetics and Phonology*. Second Edition. Malden, USA: Blackwell Publishing.

Demenko, G., Grocholewski, S., Wagner, A. & Szymański, M. (2006). Prosody Annotation for Corpus Based Speech Synthesis. In: *Proceedings of the Eleventh Australasian International Conference on Speech Science and Technology*. Auckland, New Zealand, pp. 460--465

Dutoit, T., Pagel, V., Pierret, N., Bataille, F. & van der Vrecken, O. (1996). The MBROLA Project: Towards a Set of High-Quality Speech Synthesizers Free of Use for Non-Commercial Purposes. In: *Proceedings of ICSLP 96*. Philadelphia, vol. 3, pp. 1393--1396.

Gibbon, D., Moore, R. & Winski, R. (1997). *Handbook of Standards and Resources for Spoken Language Systems*. Berlin: Mouton de Gruyter.

Gibbon, D., Mertins, I. & Moore, R. (2000). *Handbook of Multimodal and Spoken Dialogue Systems: Terminology, Resources and Product Evaluation*. New York: Kluwer Academic Publishers.

Grice, M. (2006). Intonation. In: Keith Brown, (Editor-in-Chief) *Encyclopedia of Language & Linguistics*, 2nd Edition, vol. 5. Oxford: Elsevier, pp. 778-788.

Gut, U., Milde, J-T., Voormann, H. & Heid, U. (2004). Querying Annotated Speech Corpora. In: *Proceedings of Speech Prosody 2004*. Nara, Japan, March 23-26, 2004, n.p.

Krauwer, S. (2005). ELSNET and ELRA: a common past and a common future. http://www.elda.org/article48.html, accessed 2005-10-14.

Ryalls, J. 1996. *A Basic Introduction to Speech Perception*. San Diego, California: Singular Publishing Group, Inc., and London: Singular Publishing Ltd.

Szklanny, K. & Masarek, K. (2002). PL1 - A Polish female voice for the MBROLA synthesizer. http://tcts.fpms.ac.be/synthesis/mbrola/mbrcopybin.html, accessed 2006-11-25.