

Ogmios: The UPC Text-to-Speech synthesis system for Spoken Translation

Antonio Bonafonte, Pablo D. Agüero, Jordi Adell, Javier Pérez, Asunción Moreno

Department of Signal Theory and Communication
TALP Research Center
Universitat Politècnica de Catalunya (UPC), Barcelona, Spain

Abstract

This paper presents the baseline text-to-speech system developed at UPC (Ogmios) plus our recent work on speech prosody generation and the procedures to create high quality language resources for speech synthesis. These contributions have been evaluated within the TC-STAR European project, which is focused on speech-to-speech translation. Several presented contributions have been developed in order to adapt the TTS component to the speech-to-speech translation framework. In this application, the input text is not *written-style* text but transcriptions of *talks*. Moreover, we have to cope with errors coming from the speech recognition and speech translation engines. However, in speech-to-speech translation, the source speech can be used as a valuable source of information to generate the target prosody. The general framework and first results are presented in the paper.

1 Introduction

In 2003, the TC-STAR European project started. It is envisaged as a long-term effort to advance research in all core technologies for Speech-to-Speech Translation (S2ST), such as Automatic Speech Recognition (ASR), Spoken Language Translation (SLT) and Text to Speech (TTS) synthesis. It targets the translation of European Parliament speeches and broadcast news. The languages of the project are UK-English, Castilian Spanish and Mandarin. Within the project, with regard to TTS, a long-term research goal is to produce voice in the target language providing all the information given by the source speaker. Therefore, the Text-to-Speech component is not limited to provide linguistic information, but also which is the speaker attitude toward the message, which are his/her feelings, which are the most relevant parts of the speech, etc. Furthermore, the voice should be similar to the voice of the original speaker if s/he could speak the target language.

Speech synthesis is today a consolidated research topic in speech technology. Usually, the main focus is reading text or provide information in an almost *reading* style. Speech synthesis has also been used in concept-to-speech systems, where input includes additional information which can be used to improve prosody generation. However, the use of concept-to-speech systems for very broad or unlimited domains is extremely difficult.

Generating speech in the speech-to-speech framework, in particular if statistical translation is used, is a new and challenging task for speech synthesis. On one hand, translated text is often difficult to read even for humans because it represents transcriptions of speech (not *written-style* text). The speech can include elements of what has been named *spontaneous speech*, including restarts, repetitions, corrections, fillers, etc. Even perfect recognition and translation makes speech synthesis a difficult task, and reading style is no more suitable for this task: we have to move from *reading-style* to *talking-style*. Furthermore, speech recognition and spoken translation engines introduce errors making the task even harder. In particular, recognition (and translation) of punctuation is a topic under development; and reliable punctuation is critical in speech synthesis. On

the other hand, the speech-to-speech translation framework offers a new source of information to the speech synthesis component: source speech itself. It can be used to drive the generation of synthetic speech, in particular, its prosody.

This paper presents the baseline system developed by UPC. Its architecture and an overview of the algorithms are presented in section 2. As in any unit-selection-based system, the design and the quality of the speech databases play a fundamental role. In section 3 we explain the language resources produced in order to develop the baseline system and for research on the speech-to-speech framework. Section 4 explains our approach to integrate the system in the speech-to-speech framework.

Within TC-STAR, periodic competitive evaluations are scheduled to track the advances in all speech-to-speech translation technologies. Section 5 reports the results that we have obtained in the evaluation of the prosody generation module and the speech synthesis component (whole TTS system). In order to evaluate the portability of our algorithms across languages, we have also created a new English voice with no specific research for this language. Results for this new voice are also presented here.

Finally, in section 6 we present a summary of the paper and future research directions within the project.

2 The Ogmios System

The UPC Text-to-Speech System (Ogmios) is a multilingual-focused system. As design principles, the algorithms are to some extent language-independent while language-dependencies are kept in data. In order to reduce the development cost, most of the techniques are either language independent (e.g. acoustic modules) or data driven (e.g. prosody generation, phonetic transcription). The system was originally developed in Catalan and Spanish and has been extended to work in other languages (French, Portuguese, English). The core of the system is a C++ set of modules with a common interface based on highly structured data describing linguistic relationships at different levels, ranging from a shallow description of the syntax until acoustic features of the speech segments to be concatenated. Several wrappers have been developed to

make the system compliant with some standards *de facto* (SAPI.4, SAPI.5), web servers, etc.

2.1 General description

Ogmios contains many modules, each one devoted to a specific function. All modules can be classified in three main areas: symbolic analysis, prosody generation and waveform generation, as formally defined in Pérez et al. (2006)

- **Symbolic analysis.** First, it tokenises the input text and classifies each token (punctuation, acronyms, abbreviations, cardinal and ordinal numbers, time and data expressions, Internet locators, etc.) and they are expanded into full orthographic forms. The input to this module is plain text, which can be optionally marked with the SSML language. The text is labelled with part-of-speech (POS) tags using a statistical tagger. For Spanish, shallow parsing is also added. Finally, words pronunciation is derived from a dictionary and a finite state transducer (learnt from the dictionary) predicts pronunciation of unknown words.
- **Prosody generator.** This is the principal agent in obtaining natural sounding quality of synthetic speech. There are several tasks: phrasing, f0-contour generation, segmental duration assignment and intensity contour generation. Each of these tasks are performed by a single module.
- **Waveform generation** Synthesis is performed by concatenation of recorded segments selected from a large database (see section 3). Basic units are context dependent demiphones. Acoustic and phonological features are used to select the segments. *Phrase-selection* is introduced to get all the units from phrases which are completely present in the database, and then prosody is not modified.

2.2 Prosodic modelling

Prosody generation is done by a set of modules that sequentially perform all the tasks involved in prosody modelling: phrasing, duration, intensity and intonation.

2.2.1 Phrasing

Phrasing is one of the key topics in the linguistic part of text-to-speech technologies. Phrasing consists on breaking long sentences into smaller prosodic phrases. Boundaries are acoustically characterised by a pause, a tonal change, and/or a lengthening of the last syllable. Phrase breaks have strong influence on naturalness, intelligibility and interpretation of a sentence. The presence or absence of them can produce a change in the meaning of a sentence.

In Ogmios phrasing is obtained using a CART. Each word boundary is considered a candidate to be a phrase break. The set of features used to predict phrase breaks are: punctuation marks, POS in a 5 words window (three preceding words and two following words), number of words and syllables since last phrase break, number of syllables and words until next punctuation and syntactic boundary.

2.2.2 Duration

Phone duration strongly depends on the rhythmic structure of the language. For example, English is stressed-timed while Spanish is syllable-timed. Syllable duration is predicted using a CART. Features include the structure of the syllable, represented by articulatory information of each phoneme contained in the syllable (phone identity, voicing, point, manner, vowel or consonant), the position of the syllable in the sentence and inside the intonation phrase, etc.

Once the duration of the syllable is calculated, the duration of each phoneme is obtained using a set of factors to distribute syllable duration along its phonemes. These factors are predicted using a set of features extracted from the text, such as articulatory information of the phoneme itself and the preceding and succeeding ones, position in the syllable, in the word and in the sentence, and whether the syllable is pre-pausal.

2.2.3 Intensity

The intensity of the phonemes is predicted by means of a CART. Features are again articulatory information of the actual, preceding and succeeding phone plus the position in the sentence relative to punctuation and phrase breaks.

2.2.4 Intonation

Our intonation model uses a superpositional approach combining the influence of two prosodic units: accent groups and minor phrases. Accent groups model local effects at the level of the stressed syllable and minor phrases model a long-term effect of the intonation contour. Each component of the intonation model is approximated by means of a Bèzier curve.

The intonation model is trained using JEMA: Join feature extraction and modelling approach, a new approach to train intonation models that combines the parameter extraction (step 1) and model generation (step 2) into a single loop (Agüero and Bonafonte, 2004; Agüero et al., 2004). This approach avoids requirements of continuity of fundamental frequency contours and increases parametrisation consistency. It has been successfully applied to several languages and models (Tilt, Fujisaki, Bèzier) (Rojc et al., 2005; Agüero and Bonafonte, 2005).

Parameters of the Bèzier curve are predicted using a set of features extracted from text, such as position of the prosodic unit in the sentence, number of words and syllables in the unit, position of the stressed syllable, punctuation mark, etc.

3 Language Resources

In TC-STAR, specifications to produce high-quality language resources for speech synthesis have been produced (Bonafonte et al., 2004). These specifications include corpus design, speaker selection, recording platforms and annotation. Briefly, for each baseline voice, 90K words are recorded, around 10 hours of speech. The corpus domain contain novels, parliamentary transcriptions and application words (such as numbers, dates, etc.). Three channels are recorded (96KHz, 24bits): close talk, membrane microphone, and laryngograph. For all the data, the phonetic transcription and basic prosody have been manually annotated. Furthermore, pitch labels and phonetic segmentation

of 20% of the data have been manually supervised. The speaker selection process consisted on selecting five professional speakers for each voice. Each of them recorded a set of sentences (more than twenty minutes of speech) in order to create a small synthesiser. Afterwards, a MOS test was performed. The final selection was based on several factors, as pleasantness of the voice, good articulation and also the result of the MOS. Selected speakers recorded the whole database. The TC-STAR deliverable (Bonafonte et al., 2004) also specifies language resources for intra and cross-lingual voice conversion and for supporting research on the speech-to-speech framework.

Based on these specifications, UPC have produced one male and one female baseline voices, 4 bilingual (English/Spanish) speakers for voice conversion and 4 bilingual speakers for supporting research in the speech-to-speech translation. This data will be distributed by ELDA. For the baseline voices, a preliminary phonetic segmentation was computed using the UPC speech-recognition toolkit (in the forced-alignment modality). Speaker dependent models were estimated for context-dependent demiphones. Silences were detected as side information of the training algorithm.

The likelihood provided by the HMM-forced alignment was used to select the 20% of sentences which had to be checked manually. Segmentation was also used by the prosody and phonetic labelling toolkit to ease the navigation through the files. After the phonetic transcription was supervised, the files were automatically segmented again and a pruning strategy was followed to detect problematic units (Adell et al., 2006). This caused the database to be 10% smaller but 90% of undesired units were removed. Prosodic models were trained from this database.

Pitch labels were obtained following the method described in (Pérez and Bonafonte, 2005). Pitch epochs are extracted from the laryngograph signal recorded simultaneously with the speech. We use the glottal closure instants, obtained as the instants of occurrence of the minimum of the derivative of the laryngograph signal. Since the laryngograph signal is often noisy, pitch marks are post-processed in order to obtain a cleaner estimation. They require a further correction due to the delay between the laryngograph and speech signals. We use a low-pass filtered version of the speech signal comprising only the first harmonic to locate the preceding zero-crossing point. The final pitch mark is placed on the position of the minimum before this crossing by zero.

Furthermore, 4 bilingual speakers produced data for voice conversion and for the speech-to-speech framework research. The speakers had to read one hour in each language (English and Spanish) in an expressive style. They listened to samples of the parliamentarians and then they read/interpreted each paragraph, first in one language and then in the other one.

Finally, a new voice in English has been built. No language resources were previously available in our speech synthesis system. However, we decided to build this new voice in order to test whether the architecture of our system allowed to produce new voices quite fast. First of all, a phonetic transcription lexicon was needed. We used the Unisyn Lexicon (Fitt, 2000). Then, our final state transducer was trained

from this lexicon for unknown words. Then, database was phonetically transcribed automatically, pitch marks were also automatically extracted (no manual supervision was done) by means of algorithms described above. Afterwards, the database was pruned. Prosody models are data driven, so they were learnt from the pruned segmented database.

4 Speech synthesis in S2S Translation

4.1 Prosody

In the speech-to-speech translation framework (S2ST), speech is recognised in a language (source language), then the text is translated into another language (target language) and finally synthesised. The main task of TC-STAR is translation of the European Parliament. In such application, as in translation of broadcast news or lectures, the speaker is not addressing a machine but people. So, he will not adapt his speaking style to the machine. Obviously, his speech is produced to be listened rather than to be *read*. Therefore, speech will contain a lot of information in prosody and thus, it must also be *translated* in order to transmit the whole meaning of speech.

There are two possible approaches to this problem. The first one is to make explicit models of each use of prosody; as different emotions, emphasis or speaker attitude. These models would have to be used to detect this information in the source speech and to generate the correct prosody in the target language. This approach requests for modelling emotions, also for semantic and common knowledge representation, etc. Even if these models could be widely used nowadays it would still be hard to combine them in order to generate a single prosodic model. The second approach is to model prosody *translation* implicitly, without any understanding of what it transmitted. In this paper we present an approach to translate prosodic features based on finding correspondences between features across languages. The main acoustic features related with prosody are: *intonation*, *pauses*, *duration* and *energy*. To index the information, the speech recognition word segmentation is translated mapped into the target language using the word alignments produced by the statistical machine translation. The basic idea is that acoustic features (as speaking rate and f0 characterisation) can be mapped using the alignment into the target text. Then, this acoustic-based feature is one additional feature to predict prosody in the target language. We believe that in many cases speaking rate is correlated in both languages. However, this is not imposed using a rule. The feature is used by machine-learning algorithms only if it is relevant. To train the models, we have used the four bilingual speakers (therefore, we have 4 prosodic models). The method has proven to be effective for predicting breaks and segmental duration. For intonation, the method is useful when using the bilingual speaker as source voice. However, further work is needed for using the models with a different source speaker (speaker independent models). Therefore, in this paper, the intonation of the baseline system is used. Results show that the information of the input speaker improves the expressiveness of the output.

In previous publications (Agüero et al., 2006a; Agüero et al., 2006b) we have described a proposal to generate prosody for speech-to-speech translation between Spanish

and Catalan. The basic idea in this work is to label the intonation in the *source speech*, and use these labels as an additional feature to generate the intonation of the *target speech*. The labelling will be based on acoustic measures of the pitch contour, without giving any interpretation to the movement.

The underlying hypothesis is that some paralinguistic information is correlated with the F0 contour. This has already been established in many studies related to emphasis, speaker emotions and even the audience and speaking style (see for instance Werner and Keller (1994) or Campbell (2004)). In some cases (e.g. some emotions) the F0 movements are in some extent language independent (Hirst and Cristo, 1998; Nogueiras et al., 2001). However, this is not relevant to our proposal. Our goal is not to *copy* the movements but to *label* the movements in the source speech and use this characterisation as additional features to infer the intonation model in the target language.

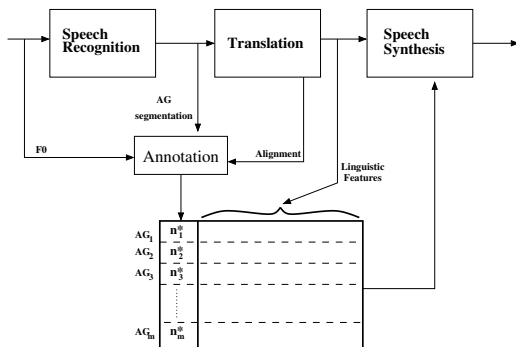


Figure 1: Annotation process uses accent group segmentation, alignment and source pitch in order to add a new feature to the vector of linguistic features for each accent group.

Figure 1 outlines the process used to generate the prosody in the complete speech-to-speech translation system. The source speech is recognised and translated into the target language. In parallel, the source speech is analysed and annotated prosodically, as described in Agüero et al. (2006a) and Agüero et al. (2006b). Based on statistical alignment, the source pitch movements are mapped to the target pitch movements. The speech synthesis component includes all the prosody generation modules. Usually these modules derive the synthetic prosody based on features extracted from text (as sentence modality, stress, number of syllables and eventually syntactic features). In this proposal these features are extended to include the new feature derived from the source speech.

4.2 Ill formed sentences

Ill-formed sentences are difficult to understand by humans, since we expect language to have a defined structure. Some studies have shown that when humans produce difficult to understand sentences they add a silent or filled pause before it (Tree, 2001; Watanabe et al., 2005). This is because speakers may need some more time to think of what they will say. At the same time, from the point of view of the listener, it indicates that what is to be said will be hard

to understand, since it is hard to produce. Based on this, we have introduced a module in order to predict pauses in places where the listener might find some difficulties to understand based on the probability of a language model. The language model is built by linear interpolation of a Part-of-Speech (POS) based n -gram and another one based on lemmas:

$$p(w_i/h_i) = \lambda_p p_p(w_i^p/h_i^p) + \lambda_l p_l(w_i^l/h_i^l) \quad (1)$$

where $p(w_i/h_i)$ is the probability of word w_i given the history of this word (h_i). p_p is the probability given by the POS n -gram and p_l the one given by the lemma n -gram. w_i^p is the POS and w_i^l the lemma of word i . λ_p and λ_l are weights for each of the models and $\lambda_p + \lambda_l = 1$.

Therefore, given a threshold p_{th} every-time $p(w_i/h_i) < p_{th}$ the sentence is considered ill-formed at this point, because of low probability. Then a break is inserted before the word w_i and intelligibility is expected to be improved by keeping listeners attention of what will be said after the break.

There is no corpus available to tune the parameters of this model so they have been set empirically. Informal tests have shown promising results. However, a formal evaluation must be performed in future work.

5 Evaluation results and discussion

The TC-STAR evaluation campaigns use different tests to assess each module: prosody modules, the TTS component, intra-lingual and cross-lingual voice conversion and expressive speech. Here we present the results that we have obtained for the prosody module and the TTS component in Spanish and English, in two evaluation campaigns (year 2004 and 2005). There is one evaluation track that includes the output of the speech recognition and translation engines. The third open evaluation campaign is scheduled for January 2007.

5.1 Prosody Module Evaluation

Evaluation of the prosody module has been done through three tests. A resynthesis-based one, and two more tests using delexicalised speech based on the work done in Sonntag and Portele (1998).

Here, the capability of the module to generate natural prosody is aimed to be evaluated. Therefore, the module input information is an XML file containing correct symbolic analysis (i.e. word normalisation, phonetic transcription, ...) and the output information is another XML file containing duration and energy for each phone, and pitch contours.

Original natural audio files are modified by means the PSOLA algorithm in order to generate new audio files with prosody generated by each of the systems, these files were ranked by judges in a 5-point scale (1: Very unnatural, 5: Very natural) and constituted the Resynthesis evaluation. Natural audio files are also given to be ranked in order to have a reference, but their pitch were moved 20Hz up and down to produce some quality degradation as PSOLA does in resynthesised audio files.

System	OQ	LE	Pr	C	A	SR	N	EL	PI	A
Spanish										
Natural	4.61	4.89	4.89	4.94	4.67	4.97	4.58	4.36	4.28	4.33
Eval. 1 Male	3.75	3.92	3.92	4.25	3.96	4.17	3.21	3.17	3.67	2.75
Eval. 1 Female	3.58	4.17	4.08	4.5	4.04	4.37	3.37	3.37	3.67	3.17
Eval. 2 Male	4.00	4.28	4.00	4.44	4.11	4.17	3.36	3.47	3.67	3.25
Eval. 2 Female	3.89	4.36	4.14	4.56	3.64	4.08	3.25	3.17	3.56	2.97
English										
Natural	4.58	4.59	4.77	4.84	4.61	4.67	4.48	4.41	4.31	4.33
Eval. 2 Female	2.84	2.92	3.02	3.49	3.25	3.83	2.26	2.13	2.82	2.28

Table 1: MOS Evaluation results of the speech synthesis component: overall speech quality (OQ), listening effort(LE), pronunciation(P), comprehension(C), articulation(A), speaking rate(SR), naturalness(N), easy of listening(EL), pleasantness(PI), and audio flow(A).

In order to perform the delexicalised tests, audio files were generated from the output XML file containing pitch contours.

Subjects were asked to say whether the delexicalised audios contained a suitable prosody for the given text. A five point scale was again used for this purposes (1:Very bad, 5:Excellent). The last test (Delex. B) consisted in choosing the most appropriate text given the delexicalised audio among 5 different ones which differ in phrase modality, boundaries, number of syllables, etc. Scores of this test go from 0 to 1 whether the text used to generate the prosody has been chosen by the subject or not. Results are presented in table 2.

System	Resynth.	Delex. A	Delex. B
Spanish			
Natural	4.19	3.58	0.75
UPC Female	N/A	3.30	0.69
UPC Male	2.48	3.25	0.61
English			
Natural	4.10	3.97	0.65
UPC Female	2.20	2.04	0.36

Table 2: Results of the prosodic module: MOS test using resynthesised sentences (Resynth), MOS test using delexicalised sentences (Delex. A) and appropriateness test based on the selection of the most appropriate paragraph to the generated delexicalised utterance (Delex. B).

Results show that natural prosody is strongly preferred with respect to synthetic. However, when working with delexicalised audio files, this difference is smaller.

5.2 Judgement of the speech synthesis component

This evaluation intends to judge the whole TTS system in several aspects that are relevant to assess an analysis of the quality of the speech synthesis. The judgement was carried out via web. Each subject was asked to rate, using 5-point scales, different aspects of the system. In Table 1 is shown the evaluation results of the speech synthesis component for several systems: systems presented in the first evaluation campaign (*Eval. 1 Male* and *Eval. 1 Female*), systems

presented in the second evaluation campaign (*Eval. 2 Male* and *Eval. 2 Female*) and natural voice.

The results show a progress with respect to previous evaluation in overall quality, listening effort, pronunciation and comprehension) for Spanish. These results are due to our work in several aspects, such as database segmentation and labelling, and an improvement of prosody modules.

We defined a different test to assess whether the use of source information increased the intelligibility of the synthetic voice. As mentioned before, the information of the source (real data from the parliamentary) was used as additional information to derive the phrasing and the segmental duration. However, for the intonation, an intonation model derived from the Spanish recordings of the bilingual speakers was used. The judges were asked if the expressiveness of the synthetic voice was appropriate in that context. Listeners agreed that the expressiveness was appropriated for both systems (using or not using information from the source speaker). Then, they were asked if the voice was non-expressive (rate=3), slightly expressive (rate=2) or very expressive (rate=1). For the female voice the results showed that the new voice is more expressive (2.64 for the baseline and 2.10 for the new version). However for the male voice the results did not show any improvement. We think that the reason is that the intonation of the bilingual male speaker was not as expressive as the voice of the baseline speaker. Improvements on phrasing and duration were shadowed by the degradation on the intonation model. This was not the case for the female speaker, where the bilingual speaker rendered very good intonation.

After the evaluation campaign, we consider that we should have included the original speech in the test. The expressiveness of the voices should be related to this reference. Furthermore, English results were good, despite no language specific work has been done for this language. The performance of the system did not decrease too much, and this supports the language-independent functionality of Ogmios.

5.3 Intelligibility with ASR+SLT

In order to evaluate the intelligibility of the system, the output of the speech recogniser plus spoken translation engines have been synthesised (500 words). The synthetic speech

has been transcribed by people. They could listen to each sample twice and they were instructed to write down exactly what they listened. The word error rate was 5.0% in Spanish and 9% in English. It can be observed how intelligibility results are correlated with the overall quality of the system.

6 Summary and further work

In this paper we have shown several aspects of the work under progress within TC-STAR project at Universitat Politècnica de Catalunya (UPC), the architecture of the Text-to-Speech system and details of prosody modules.

In addition, we have proposed how to integrate the speech synthesis system in the speech-to-speech framework. First, the prosody of the synthetic speech is predicted using acoustic features derived from the source speech. This has produced good results for phrasing and segmental duration. However, for f_0 , further work is needed to use the models in the speaker independent case. Furthermore, we have proposed a simple but effective approach to introduce pauses in the input text in order to treat ill-formed sentences originated by the speaker, Automatic Speech Recognition system (ASR) or Spoken Language Translation system (SLT). The paper reports the results of the second evaluation campaign of TC-STAR project for the prosodic modules isolated and for the whole speech synthesis component. In the case of the speech synthesis component we obtained an improvement from results of the previous evaluation campaign.

English results support that the language-independent architecture of Ogmios allows to build a voice in a new language from recorded speech in a short time (in the case of English less than one month).

7 Acknowledgements

This work has been funded by the European Union under the integrated project TC-STAR - Technology and Corpora for Speech to Speech Translation (IST-2002-FP6-506738, <http://www.tc-star.org>)

8 References

Jordi Adell, Pablo D. Agüero, and Antonio Bonafonte. 2006. Database pruning for unsupervised building of text-to-speech voice. In *Proc. of ICASSP*, May. Toulouse, France.

Pablo Daniel Agüero and Antonio Bonafonte. 2004. Intonation modeling for TTS using a joint extraction and prediction approach. *Proceedings of the International Workshop on Speech Synthesis*.

Pablo Daniel Agüero and Antonio Bonafonte. 2005. Consistent estimation of Fujisaki's intonation model parameters. *SPECOM 2005*.

Pablo Daniel Agüero, Klaus Wimmer, and Antonio Bonafonte. 2004. Joint extraction and prediction of Fujisaki's intonation model parameters. *Proceedings of International Conference on Spoken Language Processing*.

Pablo Daniel Agüero, Jordi Adell, and Antonio Bonafonte. 2006a. Prosody generation for speech-to-speech translation. In *Proceedings of ICASSP 2006*, May. Toulouse, France.

Pablo Daniel Agüero, Jordi Adell, and Antonio Bonafonte. 2006b. Prosody generation in the speech-to-speech translation framework. In *Proceedings of Speech Prosody 2006*, May. Dresden, Germany.

Antonio Bonafonte, Harald Höge, Herbert S. Töpf, Asuncion Moreno, Henk van der Heuvel, David Sünndermann, Ute Ziegenhain, Javier Pérez, and Imre Kiss. 2004. Deliverable d8: TTS - baselines and specifications. Technical report, TC-STAR project. www.tc-star.org.

Nick Campbell. 2004. Speech & expression; the value of a longitudinal corpus. In *Proceedings of LREC*. Lisbonne, Portugal.

Sue Fitt. 2000. Documentation and user guide to unisyn lexicon and post-lexical rules. Technical report, Centre for Speech Technology Research, University of Edinburgh. <http://www.cstr.ed.ac.uk/projects/unisyn/>.

Daniel Hirst and Albert Di Cristo, editors. 1998. *Intonation Systems: A review of twenty languages*. Cambridge University Press.

Albino Nogueiras, Asunción Moreno, Antonio Bonafonte, and José B. Mariño. 2001. Speech emotion recognition using Hidden Markov Models. In *EUROSPEECH 2001*, September. Aalborg, Denmark.

Javier Pérez and Antonio Bonafonte. 2005. Automatic voice-source parametrization of natural speech. In *Proc. of Interspeech 2005*, September. Lisbon, Portugal.

Javier Pérez, Antonio Bonafonte, Horst-Udo Hain, Eric Keller, Stefan Breuer, and Jilei Tian. 2006. ECESS inter-module interface specification for speech synthesis. *Proceedings of LREC Conference*.

Matej Rojc, Pablo Daniel Agüero, Antonio Bonafonte, and Zdravko Kacic. 2005. Training the Tilt intonation model using the JEMA methodology. *Eurospeech 2005*.

G. P. Sonntag and T. Portele. 1998. PURR - a method for prosody evaluation and investigation. *Journal of Computer Speech and Language, Special Issue on Evaluation in Language and Speech Technology*, 12(4):437–451, October.

Jean E. Fox Tree. 2001. Listeners' uses of *um* and *uh* in speech comprehension. *Memory & Cognition*, 29(2):320–326.

Michiko Watanabe, Keikichi Hirose, Yasuharu Den, and Nobuaki Minematsu. 2005. Filled pauses as cues to the complexity of following phrases. In *Proc. of Eurospeech*, pages 37–40, September. Lisbon, Portugal.

S. Werner and E. Keller, 1994. *Fundamentals of Speech Synthesis and Speech Recognition: Basic Concepts, State of the Art, and Future Challenges*, chapter Prosodic aspects of speech, pages 23–40. E. Keller. Chichester. John Wiley.