Chapter 8

Automatic speech recognition

Mikko Kurimo, Kalle Palomäki, Teemu Hirsimäki, Janne Pylkkönen, Ville Turunen, Sami Virpioja, Matti Varjokallio, Ulpu Remes, Heikki Kallasjoki, Reima Karhila, Teemu Ruokolainen, Tanel Alumäe, Sami Keronen, Andre Mansikkaniemi

8.1 Introduction

Automatic speech recognition (ASR) means an automated process that inputs human speech and tries to find out what was said. ASR is useful, for example, in speech-to-text applications (dictation, meeting transcription, etc.), speech-controlled interfaces, search engines for large speech or video archives, and speech-to-speech translation.

Figure 8.1 illustrates the major modules of an ASR system and their relation to applications. In *feature extraction*, signal processing techniques are applied to the speech signal in order to dig out the features that distinguish different phonemes from each other. Given the features extracted from the speech, *acoustic modeling* provides probabilities for different phonemes at different time instants. *Language modeling*, on the other hand, defines what kind of phoneme and word sequences are possible in the target language or application at hand, and what are their probabilities. The acoustic models and language models are used in *decoding* for searching the recognition hypothesis that fits best to the models. Recognition output can then be used in various applications.

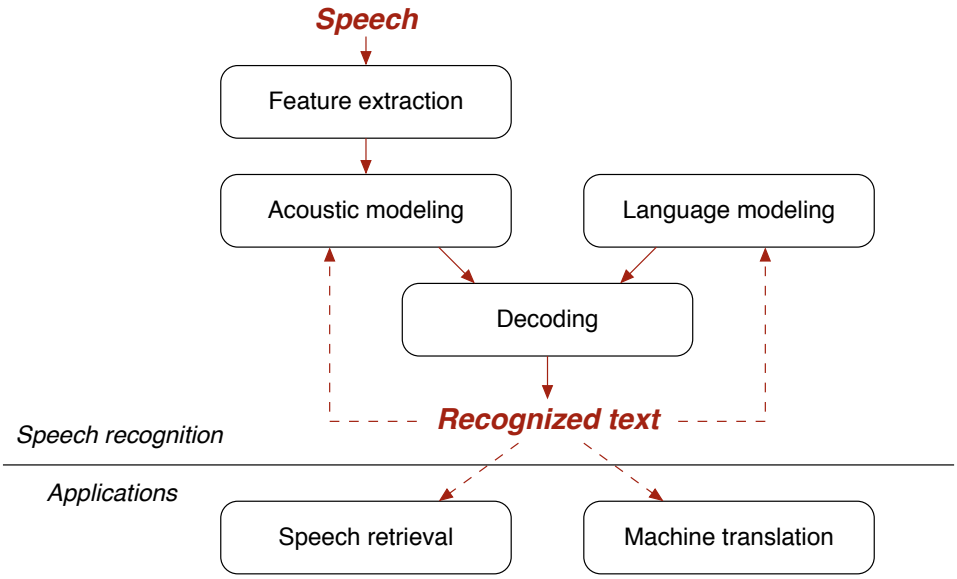

Figure 8.1: The main components of an automatic speech recognition system and their relation to speech retrieval and machine translation applications.

Our focus in ASR is large vocabulary continuous speech recognition (LVCSR). For several years, we have been developing new machine learning algorithms for each of the subfields and building a complete state-of-the-art recognizer to evaluate new methods and their impact. Originally, the recognizer was constructed for fluent and planned speech such as Finnish newsreading, where language models covering a very large vocabulary are required. Besides newsreading, other example tasks are political and academic speeches and other radio and television broadcasts where the language used is near the written style. Sofar, we have not seriously attempted to recognize Finnish spontaneous conversations, because enough Finnish training texts for learning the corresponding style do not exist. Our main training corpus for language modeling is the Finnish Language Bank at CSC. For acoustic modeling we use voice books, Finnish Broadcast Corpus at CSC and the SPEECON corpus.

In addition to the recognition of Finnish, we have performed experiments in English, Turkish and Estonian. To make this possible we have established research relations to

different top speech groups in Europe and U.S., e.g. University of Colorado, International Computer Science Institute ICSI, Stanford Research Institute SRI, IDIAP, University of Edinburgh, University of Sheffield, University of Cambridge, Bogazici University, and Tallinn University of Technology. The forms of collaboration have included researcher exchanges, special courses, workshops and joint research projects. We have also participated in several top international and national research projects funded by EU, Academy of Finland, Tekes, and our industrial partners. In the close collaboration with our Natural Language Processing group 10 we are also organizing an international competition called Morphochallenge to evaluate the best unsupervised segmentation algorithms for words into morphemes for information retrieval, statistical machine translation, LVCSR and language modeling in different languages. This challenge project is funded by EU's PASCAL network and described in Chapter 10.

In the EU FP7 project called EMIME, the aim is to develop new technologies for speech-to-speech systems. Although this has broadened the field of the group to include some aspects of speech synthesis, such as supervised and unsupervised adaptation in the same way as in ASR, text-to-speech (TTS) still plays a minor role compared to the strong ASR focus of the group.

8.2 Acoustic modeling

Acoustic modeling in automatic speech recognition means building statistical models for some meaningful speech units based on the feature vectors computed from speech. In most systems the speech signal is first chunked into overlapping 20-30 ms time windows at every 10 ms and the spectral representation is computed from each frame. A commonly used feature vector consists of mel-frequency cepstral coefficients (MFCC) which are the result of the discrete cosine transform (DCT) applied to the logarithmic mel-scaled filter bank energies. Local temporal dynamics can be captured by concatenating the first and second order delta features (time differences) to the basic feature vector.

The acoustic feature sequence in ASR is typically modeled using hidden Markov models (HMM). In basic form each phoneme is modeled by a separate HMM, where the emission distributions of the HMM states are Gaussian mixtures (GMMs). In practice, however, we need to take the phoneme context into account, so that for each phoneme there are separate HMMs for various phoneme contexts. This leads easily to very complex acoustic models where the number of parameters is in order of millions.

Estimating the parameters of complex HMM-GMM acoustic models is a very challenging task. Traditionally maximum likelihood (ML) estimation has been used, which offers simple and efficient re-estimation formulae for the parameters. However, ML estimation does not provide optimal parameter values for classification tasks such as ASR. Instead, discriminative training techniques are nowadays the state-of-the-art methods for estimating the parameters of acoustic models. They offer more detailed optimization criteria to match the estimation process with the actual recognition task. The drawback is increased computational complexity. Our implementation of the discriminative acoustic model training allows using several different training criteria such as maximum mutual information (MMI) and minimum phone error (MPE) [1]. It also enables alternative optimization methods such as gradient based optimization and constrained line search [2] in addition to the commonly used extended Baum-Welch method.

Our recent research has taken advantage of the flexibility of our system to use different discriminative training criteria by comparing different discriminative training methods in various configurations [3]. The research showed some guidelines in how to apply certain discriminative training methods in large scale acoustic model estimation.

The speech syntesis work related to the EMIME EU/FP7 project concentrates on the adaptation of HMM-based TTS models. The goal of the project is to personalize the output voice of a cross-lingual speech-to-speech system, to make it resemble the voice of the original speaker.

The features and models of TTS systems differ somewhat from those used in ASR. A shorter timestep, typically 5 ms is used, and the count of cepstral coefficients is twice or thrice that of typical ASR features. The acoustic models do not use GMMs - simple single-Gaussian models are used, but the amount of models is much higher. The TTS models are context-dependent on a more complicated level compared to the ASR models. A single phoneme has different models depending on its position within a word, syllable and sentence, as well as the surrounding phonemes.

Training acoustic models for high-quality voice for a TTS system therefore requires data of close to 1000 high-quality sentences from the target speaker. As this much data is not available in the target application of the project, the only feasible option is to train an average TTS voice and use adaptation techniques to change it to resemble the target speakers voice.

The adaptation of HMM-based TTS models is very similar to adaptation of ASR models. Maximum a posteriori (MAP) linear transformations are applied in similar fashion to

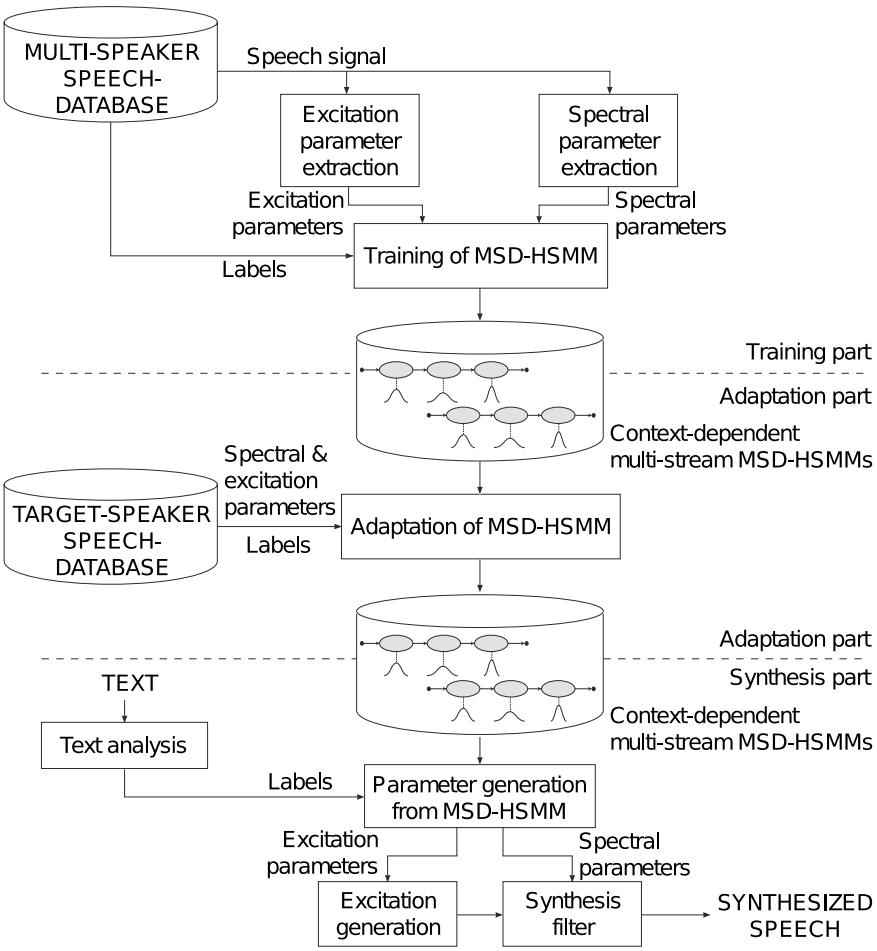

Figure 8.2: The HTS speech synthesis system for generating an average voice, adapting it to a target speaker and creating synthesized speech. From [4].

ASR adaptation. A collaborative investigation using data from several languages showed that adapting a general voice is a practical and effective way to mimic a target speaker's voice[4].

References

[1] D. Povey and P. C. Woodland, Minimum Phone Error and I-smoothing for Improved Discriminative Training. In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, Orlando, Florida, USA, pages I-105–108, 2002.

[2] P. Liu, C. Liu, H. Jiang, F. Soong, and R.-H. Wang, A constrained line search optimization method for discriminative training of hmms. In *IEEE Transactions on Audio, Speech and Language Processing*, vol. 16, no. 5, pp. 900–909, 2008.

[3] J. Pylkkönen, Investigations on Discriminative Training in Large Scale Acoustic Model Estimation. In *Proceedings of the 10th Annual Conference of the International Speech Communication Association (Interspeech 2009)*, Brighton, UK, pp. 220–223, 2009.

[4] J. Yamagishi, B. Usabaev, S. King, O. Watts, J. Dines, J. Tian,R. Hu, Y. Guan, K. Oura, K. Tokuda, R. Karhila, M. Kurimo Thousands of Voices for HMM-based Speech Synthesis. In *Proceedings of the 10th Annual Conference of the International Speech Communication Association (Interspeech 2009)*, Brighton, UK, pp. 420–423, 2009.

8.3 Language modeling

In topic adaptation of language models, we take into account the underlying topic of speech by elevating the probabilites of the subvocabulary characteristic to its topic. Via topic adaptation, we aim at improving the recognition of topically important words. The potential benefit of topic adaptation relies on the success of retrieving the underlying topic correctly. In the master's thesis [1], we discuss the topic adaptation task in relation to multimodal interfaces. In the multimodal scenario, the contextual cues with which the topic is retrieved can not be assumed reliable nor large in size. The experiments with English large vocabulary speech recognition task showed that topic adaptation with these cue assumptions is feasible. The master's thesis was conducted as a part of projects PinView and UI-ART focusing on multimodal interfaces.

For Finnish, estimating the language model probabilities for words is difficult since there is a vast number of different word forms. For example, a single verb has theoretically thousands of inflected word forms. The natural way to attack the problem is to split words into smaller fragments and build the language models on the fragments instead of whole words. Since it is not obvious how the words should be split, we have studied what kind of word fragments are optimal for speech recognition systems. Experiments in Finnish, Turkish and Estonian recognition tasks indicate that an unsupervised data-driven splitting algorithm called Morfessor (see Section 10.2) improves recognition of rare words. [2]

In speech recognition systems solutions to the problem of vocabulary growth in morphologically rich languages proposed in the literature include increasing the size of the vocabulary and segmenting words into morphs. However, in many cases, the methods have only been experimented with low-order n-gram models or compared to word-based models that do not have very large vocabularies. In [3] we study the importance of using high-order variable-length n-gram models when the language models are trained over morphs instead of whole words. Language models trained on a very large vocabulary are compared with models based on different morph segmentations. Speech recognition experiments are carried out on two highly inecting and agglutinative languages, Finnish and Estonian. The results suggest that high-order models can be essential in morph-based speech recognition, even when lattices are generated for two-pass recognition. The analysis of recognition errors [4] reveal that the high-order morph language models improve especially the recognition of previously unseen words.

References

[1] T. Ruokolainen. Topic adaptation for speech recognition in multimodal environment. *Master's thesis, Helsinki University of Technology*, 2009.

[2] E. Arisoy, M. Kurimo, M. Saraclar, T. Hirsimäki, J. Pylkkönen, T. Alumäe and H. Sak. Statistical Language Modeling for Automatic Speech Recognition of Agglutinative Languages, *Speech Recognition*, pages 193–204. I-Tech, Vienna, Austria, 2008.

[3] T. Hirsimäki, J. Pylkkönen and M. Kurimo. Importance of high-order n-gram models in morph-based speech recognition. *IEEE Transactions on Audio, Speech and Language Processing*, 17(4):724–732, May 2009.

[4] T. Hirsimäki and M. Kurimo. Analysing recognition errors in unlimited-vocabulary speech recognition. In *Proceedings of the 2009 Annual Conference of the North ameri-*

can Chapter of the association for Computational Linguistics, NAACL 2009, Boulder, CO, May 31 – June 5 2009.

8.4 Applications and tasks

Speech retrieval and indexing

Large amounts of information is produced in spoken form. In addition to TV and radio broadcasts, more and more material is distributed on the Internet in the form of podcasts and video sharing web sites. There is an increasing need for content based retrieval of this material. Speech retrieval systems consist of two parts. First, an automatic speech recognition system is used to transcribe the speech into textual form. Second, an index is built based on this information.

The vocabulary of the speech recognizer limits the possible words that can be retrieved. Any word that is not in the vocabulary will not be recognized correctly and thus can not be used in retrieval. This is especially problematic since the rare words, such as proper names, that may not be in the vocabulary are often the most interesting from retrieval point of view. Our speech retrieval system addresses this problem by using morpheme-like units produced by the Morfessor algorithm. Any word in speech can now potentially be recognized by recognizing its component morphemes. The recognizer transcribes the text as a string of morpheme-like units and these units can also be used as index terms. We have shown that the morph-based approach for speech search suffers significantly less from OOV query words than a word based method [1].

Retrieval performance was further improved by utilizing alternative recognition candidates from the recognizer [1]. Speech recognizers typically produce only the most likely string of words, the 1-best hypothesis. Retrieval performance is decreased if a relevant term is misrecognized and is thus missing from the transcript. However, it is possible that the correct term was considered by the recognizer but was not the top choice. Thus, retrieval performance can be improved by extracting these alternative results from the recognizer and adding them to the index. A *confusion network* [2] provides a convenient representation of the competing terms along with a probability value for each term.

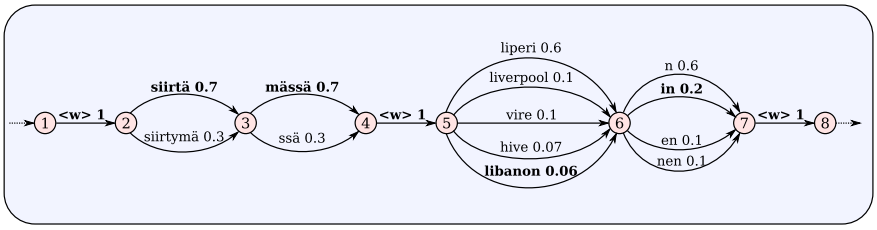

Figure 8.3: A confusion network of alternative recognition candidates for a segment of speech. `<w>` marks a word break boundary. The correct morphs are in bold.

Speech-to-speech translation

Speech-to-speech machine translation is in some ways the peak of natural language processing, in that it deals directly with our (humans') original, oral mode of communication (as opposed to derived written language). As such, it presents several important challenges:

1. Automatic speech recognition of the input using state-of-the-art acoustic and language modeling, adaptation and decoding

2. Statistical machine translation of either the recognized most likely speech transcript or the confusion network or the whole lattice including all the best hypothesis

3. Speech synthesis to turn the translation output into intelligible speech using the state-of-the-art synthesis models and adaptation

4. Intergration of all these components to aim at the best possible output and tolerate errors that may happen in each phase

A pilot study of Finnish-English speech-to-speech translation was carried out in the lab as a joint effort of the Speech Recognition, Natural Language Processing (Ch. 10) and Computational Cognitive Systems (Ch. 11) groups [3]. The domain selected for our experiments was heavily influenced by the available bilingual (Finnish and English) and bimodal (text and speech) data. Because none is readily yet available, we put one together using the Bible. As the first approach we utilized the existing components, and tried to weave them together in an optimal way. To recognize speech into word sequences we applied our morpheme-based unlimited vocabulary continuous speech recognizer [4]. As a Finnish acoustic model the system utilized multi-speaker hidden Markov models with Gaussian mixtures of mel-cepstral input features for state-tied cross-word triphones. The statistical language model was trained using our growing varigram model [5] with unsupervised morpheme-like units derived from Morfessor Baseline [6]. In addition to the Bible the training data included texts from various sources including newspapers, books and newswire stories totally about 150 million words. For translation, we trained the Moses system [7] on the same word and morpheme units as utilized in the language modeling units of our speech recognizer. For speech synthesis, we used Festival [8], including the built-in English voice and a Finnish voice developed at University of Helsinki. Further research on statistical machine translation is described in Section 13.

References

[1] V.T. Turunen. Reducing the effect of OOV query words by using morph-based spoken document retrieval. *Proceedings of the 9th Annual Conference of the International Speech Communication Association (Interspeech 2008)*, pages 2158–2161, September 2008.

[2] L. Mangu, E. Brill, and A. Stolcke. Finding consensus in speech recognition: word error minimization and other applications of confusion networks. *Computer Speech And Language*, 14:373–400, 2000.

[3] David Ellis, Mathias Creutz, Timo Honkela, and Mikko Kurimo. Speech to speech machine translation: Biblical chatter from Finnish to English. In *Proceedings of the IJCNLP-08 Workshop on NLP for Less Privileged Languages*, pages 123-130, Hyderabad, India, January 2008. Asian Federation of Natural Language Processing.

[4] Teemu Hirsimäki, Mathias Creutz, Vesa Siivola, Mikko Kurimo, Sami Virpioja and Janne Pylkkönen 2006. Unlimited vocabulary speech recognition with morph language models applied to Finnish. *Computer Speech & Language* 20(4):515–541.

[5] Vesa Siivola Language models for automatic speech recognition: construction and complexity control. Doctoral thesis, Dissertations in Computer and Information Science, Report D21, Helsinki University of Technology, Espoo, Finland, 2006.

[6] Mathias Creutz and Krista Lagus. Unsupervised models for morpheme segmentation and morphology learning. *ACM Transactions on Speech and Language Processing*, Volume 4, Issue 1, Article 3, January 2007.

[7] Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Ondrej Bojar, Alexandra Constantin, and Evan Herb. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the ACL 2007 Demo and Poster Sessions*, pages 177–180, 2007.

[8] The Festival Speech Synthesis System. University of Edinburgh. *http://festvox.org*

8.5 Noise robust speech recognition

Missing feature approaches

Using missing feature methods for noise compensation in automatic speech recognition is based on partitioning the compressed spectrographic representation of a noise corrupted speech signal to speech dominated i.e. realiable regions and noise dominated i.e. unreliable regions as illustrated in Figure 8.4. Information in the unreliable regions is assumed missing, so either the speech recognition system should ignore the unreliable components or the missing values be reconstructed using e.g. cluster-based imputation [1]. Experiments reported in [2] suggested that cluster-based imputation can significantly improve LVCSR performance under environmental noise but may not fully allow for simultaneous speaker or environment-based adaptation. We therefore modified the method to account for acoustic model adaptation estimated prior to reconstruction, which improved the speech recognition performance in certain noise environments as discussed in [3]. Additionally, we have been developing missing feature techniques that are particularly robust in the presence of reverberation noise [4, 5] and models that mimic certain principles of human speech recognition especially in the binaural system [6].

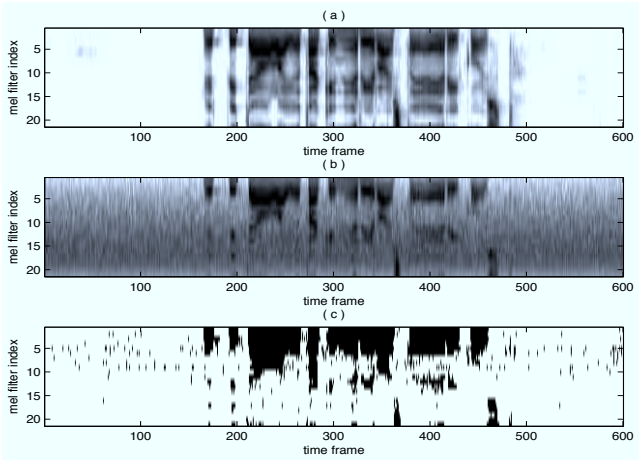

Figure 8.4: Logarithmic mel spectrogram of (a) an utterance recorded in quiet environment and (b) the same utterance corrupted with additive noise. The noise mask (c) constructed for the noisy speech signal indicates the speech dominated regions in black and the noise dominated regions in white.

Noise robust feature representations

One approach to noise robust speech recognition is to search for feature representations that are less affected by changes in environmental noise. In particular, common feature extraction schemes based on the short-time spectrum of the speech signal can be made more robust by using an estimate of the spectral envelope instead.

The stabilised weighted linear prediction (SWLP) signal modeling method [7], recently developed at the Department of Signal Processing and Acoustics at Helsinki University of Technology, was used to implement a more robust variant of the MFCC features currently used by our speech recognition system. The performance of the new features was evaluated in different noisy real-world environments using the SPEECON corpus. Improvements in

recognition rates were found in the case where acoustic models trained using clean speech only were used to recognize speech corrupted by noise [8, 9].

References

[1] B. Raj and R. M. Stern, Missing-feature approaches in speech recognition. *IEEE Signal Processing Magazine*, vol. 22, pages 101–116, 2005.

[2] U. Remes, K. J. Palomäki, and M. Kurimo, Missing feature reconstruction and acoustic model adaptation combined for large vocabulary continuous speech recognition. In *Proceedings of the 16th European Signal Processing Conference*, Lausanne, Switzerland, 2008.

[3] U. Remes, K. J. Palomäki, and M. Kurimo, Robust automatic speech recognition using acoustic model adaptation prior to missing feature reconstruction. In *Proceedings of the 17th European Signal Processing Conference*, Glasgow, Scotland, UK, pages 535–539, 2009.

[4] K. J. Palomäki, G. J. Brown and J. Barker, Techniques for handling convolutional distortion with "missing data" automatic speech recognition. *Speech Communication*, vol. 43, pages 123-142, 2004.

[5] G. J. Brown and K. J. Palomäki, A reverberation-robust automatic speech recognition system based on temporal masking (Abstract). *J. Acoust. Soc. Am.*, vol. 123, Acoustics 2008, Paris, France, page 2978, 2008.

[6] K. J. Palomäki and G. J. Brown, A computational model of binaural speech intelligibility level difference (Abstract). *J. Acoust. Soc. Am.*, vol 123, Acoustics 2008, Paris, France, page 3715, 2008.

[7] C. Magi, J. Pohjalainen, T. Bäckström, and P. Alku, Stabilised weighted linear prediction. *Speech Communication*, volume 51, issue 5, pages 401–411, 2009.

[8] H. Kallasjoki, K. J. Palomäki, C. Magi, P. Alku, and M. Kurimo, Noise robust LVCSR feature extraction based on stabilized weighted linear prediction. In *Proceedings of the 13th International Conference Speech and Computer*, St. Petersburg, Russia, pages 221–225, 2009.

[9] J. Pohjalainen, H. Kallasjoki, P. Alku, K. J. Palomäki, and M. Kurimo, Weighted linear prediction for speech analysis in noisy conditions. In *Proceedings of the 10th Annual Conference of the International Speech Communication Association*, Brighton, UK, pages 1315–1318, 2009.