Chapter 8

Automatic speech recognition

Mikko Kurimo, Kalle Palomäki, Vesa Siivola, Teemu Hirsimäki, Janne Pylkkönen, Ville Turunen, Sami Virpioja, Matti Varjokallio, Ulpu Remes, Antti Puurula

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.

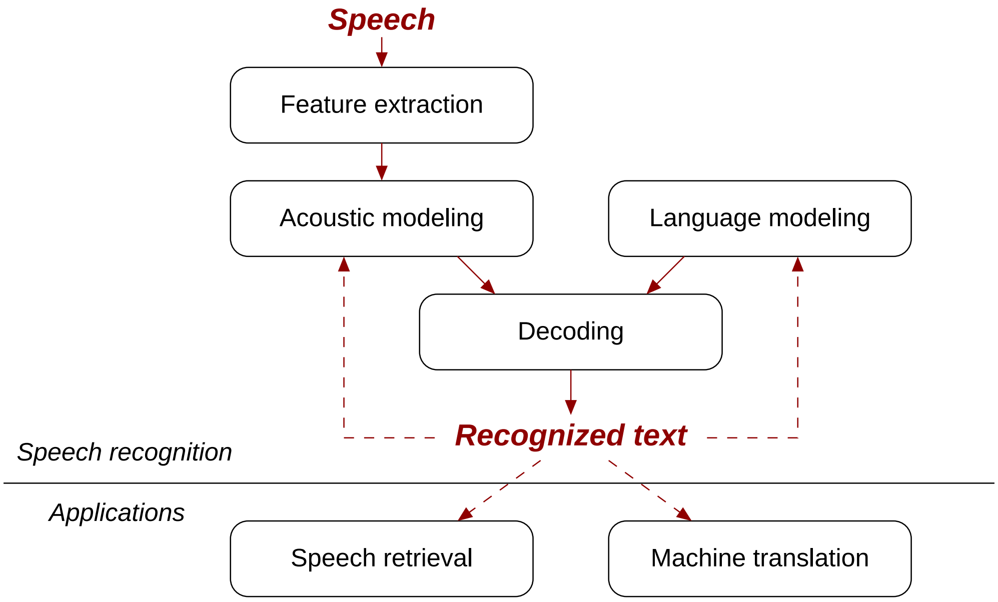

Figure 8.1: The main components of an automatic speech recognition system and their relation to speech retrieval and machine translation applications.

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, 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, LVCSR and language modeling in different languages. This challenge project is funded by EU's PASCAL network and described in Chapter 10.

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.

Although the use of DCT and delta features are well-established methods for processing speech features, they are by no means optimal. Better features can be constructed by learning from the data which features would best discriminate between speech sounds. A well known method for this is the linear discriminant analysis (LDA), which can be used to process the spectral input for creating new discriminative features. As a simple method LDA has its limitations, and therefore in [1] we studied different methods to enhance its operation. The result was the pairwise linear discriminant (PLD) features, which unlike most LDA extensions are simple to compute but still work in speech recognition better than the traditional methods.

Closely connected to the feature extraction is the speaker-wise normalization of the features. One commonly used method is the vocal tract length normalization (VTLN). It requires estimating only a single normalization parameter yet still provides significant improvements to the speech recognition. The estimation, however, can not be done in closed form, so an exhaustive search over a range of parameters is usually used. We have devised a method which greatly simplifies the estimation of the VTLN parameter but still gives competitive performance [2]. It is especially attractive when used with discriminative feature extraction, such as with PLD.

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. An example is shown in Figure 8.2. 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.

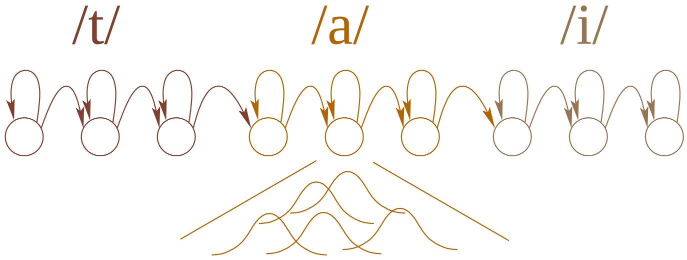

Figure 8.2: Each phoneme is modeled with a hidden Markov model, usually consisting of three states. The state distributions are modeled by Gaussian mixture models.

To limit the number of parameters and thereby allow robust estimation of the acoustic models, the covariance matrices of the Gaussian mixture components are usually assumed diagonal. This is a relatively reasonable assumption, because there is typically a whitening transform (DCT or similar) applied to the feature vector. The uncorrelatedness is, however, a global property and there are always correlations on the state level. The correlations can be modeled by adding more mixture components in the direction of most variance, which is sometimes called as *implicit covariance modeling.* Modeling covariances *explicitly* instead has some clear benefits as fewer modeling assumptions typically lead to more robust models. Constraining the exponential parameters of the Gaussians to a subspace is appealing for speech recognition, as the computational cost of the acoustic model is also decreased. A subspace constraint on the inverse covariance matrices was shown to give a good performance [3] for LVCSR tasks.

To ensure high quality research we constantly put considerable effort to keep our speech recognition system up-to-date. One major recent improvement to our system has been the introduction of discriminative acoustic training. The use of discriminative training has been a growing trend during the last decade and some form of it is now a necessity for a state-of-the-art system. Our implementation allows using several different training criteria such as maximum mutual information (MMI) and minimum phone error (MPE) [4] over the traditional maximum likelihood (ML) training. It also enables gradient based optimization in addition to the commonly used extended Baum-Welch method. Discriminative training techniques have already given very promising results and they will be an important research direction in the future.

Speaker segmentation

In addition to feature normalization methods such as the vocal tract length normalization (VTLN), acoustic model adaptation is often used for increased robustness against speaker variation. Speaker normalization and adaptation generally improve the speech recognition performance substantially, but they cannot be applied unless the speech recognition system knows who spoke and when. Often there is no such information about the speakers, but automatic speaker segmentation is needed. Speaker segmentation (i) divides the audio to speaker turns (speaker change detection) and (ii) labels the turns according to speaker (speaker tracking) as illustrated in Figure 8.3. While most speaker segmentation methods have been developed primarily for audio content or spoken dialogue analysis, we focused on speaker segmentation for speaker adaptation. We developed a speaker tracking method that seeks to directly maximize the feature likelihood when we assume the features are adapted to speaker using the segmentation results and acoustic model adaptation with constrained maximum likelihood linear regression (CMLLR). The proposed method performed well when tested on Finnish television news audio in [5].

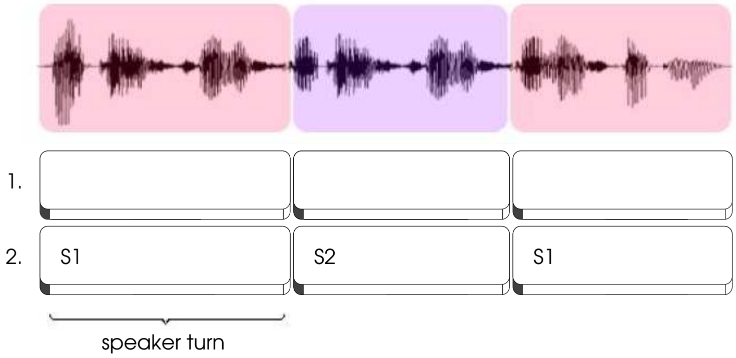

Figure 8.3: Speaker segmentation first divides the audio to speaker turns according to where speakers change and then labels the detected turns. Speaker labels are created on-line and no prior information about the speakers (e.g. training data or speaker models) is needed.

Recognition of reverberant speech

Research in the acoustic modeling for large vocabulary continuous speech recognition was concentrated mostly on fairly noise free conditions (see Sect. 8.2). In the field noise robust speech recognition we have been developing techniques suitable for recognition in highly reverberant spaces. This research has been collaborative with the University of Sheffield. Our approach is based on missing data approach [9], in which noisy, reverberated regions are treated as unreliable and noise free regions as reliable evidence of speech. Different treatments of reliable and unreliable parts of speech is achieved by a modification of Gaussian mixture model proposed by Cooke et al. [9]. Our approach to reverberant speech recognition is based on detecting reliable regions of speech from strong onsets at modulation rates characteristic to speech [8]. In recent developments of the model we have sought modeling solutions that more closely match on perceptual data considering the recognition of reverberant speech by human listeners [6, 7].

References

[1] J. Pylkkönen, LDA Based Feature Estimation Methods for LVCSR. In *Proceedings of the 9th International Conference on Spoken Language Processing (Interspeech 2006)*, Pittsburgh PA, USA, pages 389–392, 2006.

[2] J. Pylkkönen, Estimating VTLN Warping Factors by Distribution Matching. In *Proceedings of the 8th Annual Conference of the International Speech Communication Association (Interspeech 2007)*, Antwerp, Belgium, pages 270–273, 2007.

[3] M. Varjokallio, M. Kurimo, Comparison of Subspace Methods for Gaussian Mixture Models in Automatic Speech Recognition. In *Proceedings of the 8th Annual Conference of the International Speech Communication Association (Interspeech 2007)*, Antwerp, Belgium, pages 2121-2124, 2007.

[4] 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.

[5] U. Remes, J. Pylkkönen, and M. Kurimo, Segregation of Speakers for Speaker Adaptation in TV News Audio. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, Honolulu, Hawaii, USA, pages IV-481–484, 2007.

[6] G. J. Brown and K. J. Palomäki Reverberation, in *Computational Auditory Scene Analysis: Principles, Algorithms and Applications*, eds. by DeLiang Wang and Guy J. Brown, Wiley/IEEE Press, 2006.

[7] G. J. Brown and K. J. Palomäki A reverberation-robust automatic speech recognition system based on temporal masking, Research abstract accepted to Acoustics 2008, Paris, France.

[8] K. J. Palomäki, G. J. Brown and J. Barker, Recognition of reverberant speech using full cepstral features and spectral missing data,*Proceedings the IEEE International Conference on Acoustics, Speech, and Signal Processing, Tolouse, France, vol. 1, 289-292, 2006.*

[9] M.P. Cooke, P. Green, L. Josifovski, and A. Vizinho, Robust automatic speech recognition with missing and unreliable acoustic data,*Speech Comm.*, vol. 34, pp. 267 285, 2001.

8.3 Language modeling

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 10.1) improves recognition of rare words. [1]

N-gram models are the most widely used language models in large vocabulary continuous speech recognition. Since the size of the model grows rapidly with respect to the model order and available training data, many methods have been proposed for pruning the least relevant n-grams from the model. However, correct smoothing of the n-gram probability distributions is important and performance may degrade significantly if pruning conflicts with smoothing. In the journal paper [2] we show that some of the commonly used pruning methods do not take into account how removing an n-gram should modify the backoff distributions in the state-of-the-art Kneser-Ney smoothing. We also present two new algorithms: one for pruning Kneser-Ney smoothed models, and one for growing them incrementally. Experiments on Finnish and English text corpora show that the proposed pruning algorithm provides considerable improvements over previous pruning algorithms on Kneser-Ney smoothed models and is also better than the baseline entropy pruned Good-Turing smoothed models.

Representing the language model compactly is important in recognition systems targeted for small devices with limited memory resources. In [3], we have extended the compressed language model structure proposed earlier in the literature. By separating n-grams that are prefixes to longer n-grams, redundant information can be omitted. Experiments on English 4-gram models and Finnish 6-gram models show that extended structure can achieve up to 30 % lossless memory reductions when compared to the baseline structure.

Another common method for decreasing the size of the n-gram models is clustering of the model units. However, if size of the lexicon is very small, as in models based on statistical morpheme-like units (see, e.g., [1]), clustering of individual units is not so useful. Instead, we have studied how sequences of the morpheme-like units can be clustered to achieve improvements in speech recognition. When the clustered sequences are histories (context parts) of the n-grams, it is easy to combine the clustering to the incremental growing of the model applied in, e.g., [2]. Maximum a posteriori estimation can be used to make a compromise between the model size and accuracy. The experiments show that the clustering is useful especially if very compact models are required. [4]

References

[1] M. Creutz, T. Hirsimäki, M. Kurimo, A. Puurula, J. Pylkkönen, V. Siivola, M. Varjokallio, E. Arisoy, M. Saraçlar, and A. Stolcke. Morph-based speech recognition and modeling of out-of-vocabulary words across languages. *ACM Transactions on Speech and Language Processing*, 5(1), 2007.

[2] V. Siivola, T. Hirsimäki, and S. Virpioja. On Growing and Pruning Kneser-Ney Smoothed N-Gram Models. *IEEE Transactions on Audio, Speech and Language Processing*, 15(5), pages 1617–1624, 2007.

[3] T. Hirsimäki. On Compressing N-gram Language Models. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, Honolulu, Hawaii, USA, pages IV-949–952, 2007.

[4] S. Virpioja and M. Kurimo. Compact N-gram Models by Incremental Growing and Clustering of Histories. In *Proceedings of the 9th International Conference on Spoken Language Processing (Interspeech 2006)*, Pittsburgh PA, USA, pages 1037–1040, 2006.

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 as illustrated in Figure 8.4. 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.

One problem of using morpheme-like units as index terms is that different inflected forms of the same word can produce different stems when they are split to morphemes. However, we would like to retrieve the speech document no matter what inflected form of the word is used. This resembles the problem of synonyms. We have countered this problem by applying Latent Semantic Indexing to the morpheme-based retrieval approach [1]. The method projects different stems of the same word to the same dimension that represents the true, latent, meaning of the term.

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. However, as most terms in the network were in fact not spoken, the indexing method must be designed so that it is not degraded by these spurious terms. In [3], we compare methods that use the probability and rank of the terms to weigh the index terms properly and show improved performance of the retrieval system.

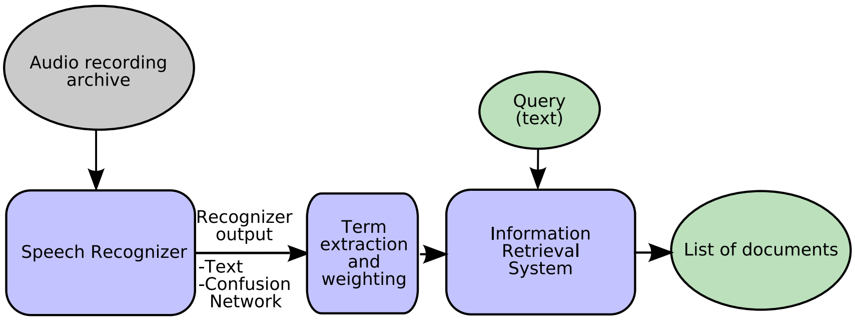

Figure 8.4: Overview of a spoken document retrieval system.

Estonian speech recognition

For agglutinative languages, like Finnish, Estonian, Hungarian and Turkish, it is practically impossible to build a word-based lexicon for speech recognition that would cover all the relevant words. The problem is that words are generally formed by concatenating several prefixes and suffixes to the word roots. Together with compounding and inflections this leads to millions of different, but still frequent word forms that can not be trivially split into meaningful parts. For some languages there exists rule-based morphological analyzers that can perform this splitting, but they are laborious to create and due to the handcrafted rules, they also suffer from an out-of-vocabulary problem.

In a pilot study of language and task portability of our speech recognition and language modeling tools, we created an Estonian speech recognizer. The text corpus used to learn the morph units and train the statistical language model consisted of newspapers and books, altogether about 55 million words [4]. The speech corpus consisted of over 200 hours and 1300 speakers, recorded from telephone [5], i.e. 8 kHz sampling rate and narrow band data instead of 16 kHz and normal (full) bandwidth that we have used for Finnish data. The speaker independence, together with the telephone quality and occasional background noises, made this task more difficult than our Finnish ones, but with the help of our learning and adaptive models we were still able to reach good recognition results and demonstrate a performance that was superior to the word-based reference systems [6, 7].

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 10 and Computational Cognitive Systems 10.3 groups. 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 [8]. 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 [9] with unsupervised morpheme-like units derived from Morfessor Baseline [10]. 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 [11] on the same word and morpheme units as utilized in the language modeling units of our speech recognizer. For speech synthesis, we used Festival [12], including the built-in English voice and a Finnish voice developed at University of Helsinki.

References

[1] V. Turunen and M. Kurimo Using Latent Semantic Indexing for Morph-based Spoken Document Retrieval, *Proceedings of the 9th International Conference on Spoken Language Processing (Interspeech 2006)*, Pittsburgh PA, USA, pages 389–392, 2006.

[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] V. Turunen and M. Kurimo Indexing Confusion Networks for Morph-based Spoken Document Retrieval, *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and development in information retrieval*, Amsterdam, The Netherlands, pages 631–638, 2007.

[4] Segakorpus. 2005. Segakorpus - Mixed Corpus of Estonian. Tartu University. *http://test.cl.ut.ee/korpused/*.

[5] Einar Meister, Jürgen Lasn and Lya Meister 2002. Estonian SpeechDat: a project in progress. In *Proceedings of the Fonetiikan Päivät - Phonetics Symposium 2002 in Finland*, 21–26.

[6] Mikko Kurimo, Antti Puurula, Ebru Arisoy, Vesa Siivola, Teemu Hirsimäki, Janne Pylkkönen, Tanel Alumae and Murat Saraclar 2006. Unlimited vocabulary speech recognition for agglutinative languages. In *Human Language Technology, Conference of the North American Chapter of the Association for Computational Linguistics.* HLT-NAACL 2006. New York, USA

[7] Antti Puurula and Mikko Kurimo 2007. Vocabulary Decomposition for Estonian Open Vocabulary Speech Recognition. In *Proceedings of ACL 2007.*

[8] 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.

[9] 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.

[10] Mathias Creutz. Induction of the Morphology of Natural Language: Unsupervised Morpheme Segmentation with Application to Automatic Speech Recognition. Doctoral thesis, Dissertations in Computer and Information Science, Report D13, Helsinki University of Technology, Espoo, Finland, 2006.

[11] 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.

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