Chapter 8

Automatic speech recognition

Mikko Kurimo, Kalle Palomäki, Janne Pylkkönen, Ville T. Turunen, Sami Virpioja, Ulpu Remes, Heikki Kallasjoki, Reima Karhila, Teemu Ruokolainen, Tanel Alumäe, Sami Keronen, André Mansikkaniemi, Peter Smit, Rama Sanand Doddipatla, Seppo Enarvi

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.

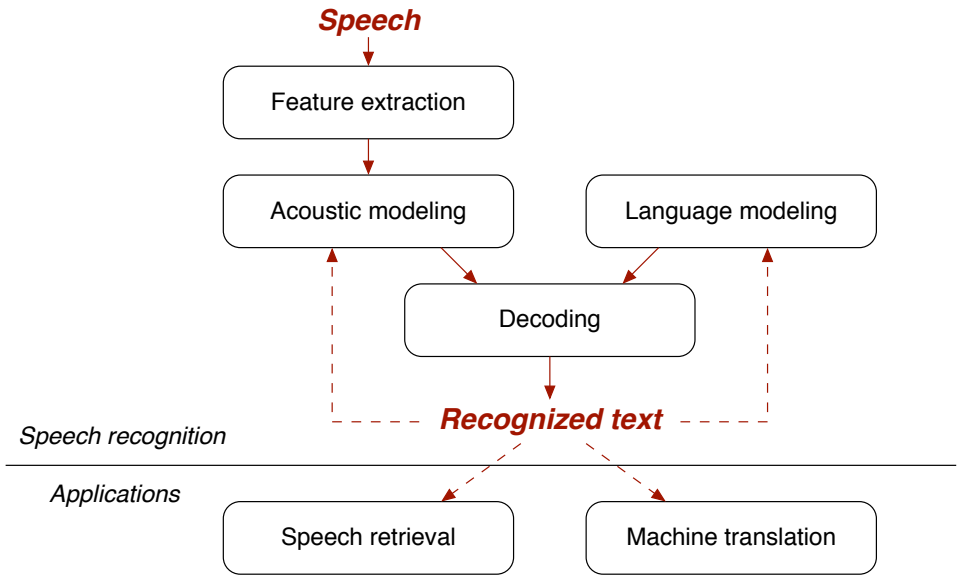

Figure 8.1: The main components of an automatic speech recognition system and their relation to speech retrieval and machine translation applications.

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, Tallinn University of Technology, and Nagoya Institute 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 Morpho Challenge 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 2008-2011, the aim was to develop new technologies for spoken multilingual integration, such as speech-to-speech translation systems. This has broadened the field of the group to include some aspects of text-to-speech synthesis (TTS), such as supervised and unsupervised adaptation in the same way as in ASR. Successors of this project include a new EU FP7 project Simple4All which aims at developing unsupervised machine learning tools for rapid data-driven development for new TTS systems by adaptation and a new project Perso which aims at developing new Finnish TTS systems by adaptation.

Other new openings in the group are developing adaptation methods for special purpose dictation (e.g. in medical domain in Mobster project), using ASR in various multimodal human-computer interaction (e.g. in augmented reality in UI-ART project), and audiovisual indexing (e.g. television broadcasts in NextMedia project).

8.2 Training and adaptation of acoustic models

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 a simple system 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. In that case each phoneme is modeled by multiple HMMs, representing different neighboring phonemes. 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]. Also alternative optimization methods such as gradient based optimization and constrained line search [2] can be used in addition to the commonly used extended Baum-Welch method. Our recent research has concentrated on comparing the different optimization strategies and finding the most effective ways to train well-performing robust acoustic models [3].

As acoustic models have a vast amount of parameters, a substantial amount of data is needed to train these models robustly. In the case a model needs to be targeted to a specific speaker, speaker group or other condition, not always sufficient data is available. The generic solution for this is to use adaptation methods like Constrained Maximum Likelihood Linear Regression [4] to transform a generic model in to a specific model using a limited amount of data. In [5] and [6] this method was repeatedly applied to a model, so that first a transformation to a foreign accented model was made and successively a transformation to a speaker-specific model. These stacked transformations improved up to 30% recognition accuracy, depending on the accent and amount of available data for the speaker. In Figure 8.2 the improvement in word error rate is shown for different amounts of speaker adaptation data and for both a native and a mixed acoustic model.

Vocal Tract Length Normalization (VTLN) has become an integral part of the standard adaptation toolkit for ASR. This method approximates physical properties of each speaker's vocal tract and shifts accordingly the frequency components of the speech to be recognized. The simple old school way of applying VTLN was to warp the cut-off frequencies in the filter bank analysis, before transforming the frequency channels of the

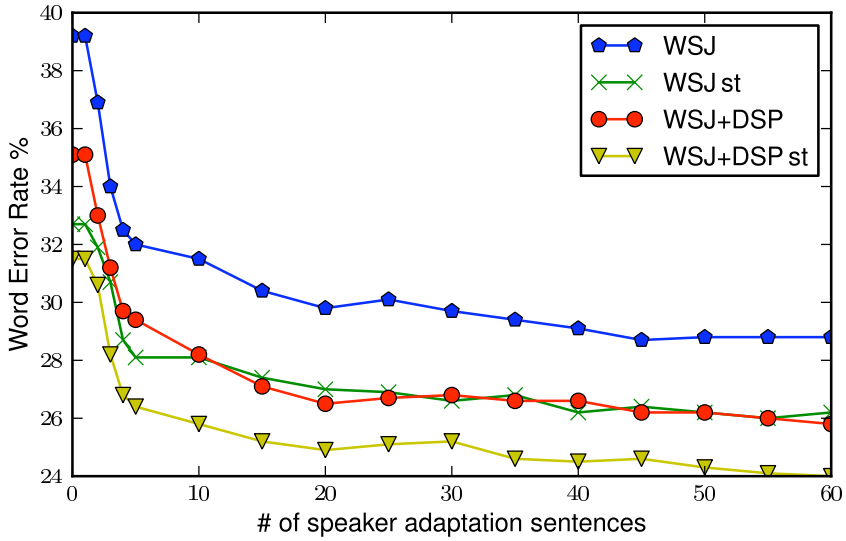

Figure 8.2: This figure shows the improvement that Stacked transformations (*st*) give over normal CMLLR adaptation. The WSJ is native English and the DSP dataset is Finnish-accented English speech. Stacked transformation have the most effect when only a small number of adaptation sentences is used.

speech sample to cepstral components. In the current approach, VTLN is represented as a CMLLR-style linear transformation on the conventional MFCC features. Using VTLN as a linear transformation on the MFCC features allowed us to study the curious interplay of CMLLR and VTLN adaptation methods and the use of VTLN to to boost other speaker adaptation methods [7].

Acoustic modeling of parametric speech synthesis

The rising paradigm of HMM-based statistical parametric speech synthesis relies on ASR-style acoustic modelling. Speech synthesis, or Text-To-Speech (TTS) models are more descriptive and less generalized than the ASR models. They try to accurately describe the numerous, variously stressed phones, and therefore the model sets are much larger than the ASR model sets. Training acoustic models for high-quality voice for a TTS system requires data of close to 1000 high-quality sentences from the target speaker. 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 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[8].

The speech synthesis work related to the EMIME EU/FP7 project concentrated on the adaptation of HMM-based TTS models. The goal of the project was to personalize the output voice of a cross-lingual speech-to-speech system, to make it resemble the voice of the original speaker [9]. This is accomplished by adapting the acoustic features of the synthesis model set in one language (Source language, L1) and mapping these transformations to a second model set (Target language, L2). The goal of the Cross-Lingual Speaker Adaptation (CLSA) is to effectively model speakers' speech in another language. As a

person's speech in a foreign language depends, beside physical characteristics, also very much on the environmental factors - mostly how much and in what kind of linguistic environment has the speaker practised speaking the language, it is almost impossible to predict how a person would in reality sound in the second language. We investigated what kind of expectations listeners usually have about a speaker's voice in a second language, and particularly whether the listeners preferred a foreign- or native accented voice model for basis of adaptation, a very important aspect in real-life situation where only little data is available for adaptation [10].

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] M.J.F. Gales, Maximum likelihood linear transformations for HMM-based speech recognition. In *Computer speech and language*, vol. 12, pp. 75–98, 1998.

[5] P. Smit and M. Kurimo, Using stacked transformations for recognizing foreign accented speech. In *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*, pp. 5008–5011, May 2011.

[6] P. Smit, Stacked transformations for foreign accented speech recognition. *Masters' thesis*

[7] D.R. Sanand and M. Kurimo, A Study on Combining VTLN and SAT to Improve the Performance of Automatic Speech Recognition. In *Proceedings of the 12th Annual Conference of the International Speech Communication Association, INTERSPEECH*, Florence, August 2011.

[8] 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.

[9] M. Wester, J. Dines, M. Gibson, H. Liang, Y.-J. Wu, L. Saheer, S. King, K. Oura, P. N. Garner, W. Byrne, Y. Guan, T. Hirsimäki, R. Karhila, M. Kurimo, M. Shannon, S. Shiota, J. Tian, K. Tokuda, and J. Yamagishi, Speaker adaptation and the evaluation of speaker similarity in the EMIME speech-to-speech translation project. In *Proceedings of the 7th ISCA Speech Synthesis Workshop, SSW7*, ISCA, September 2010.

[10] R. Karhila and M. Wester, Rapid adaptation of foreign-accented HMM-based speech synthesis. In *Proceedings of the 12th Annual Conference of the International Speech Communication Association, INTERSPEECH*, Florence, August 2011.

8.3 Noise robust speech recognition

Despite the steady progress in speech technology, robustness to background noise remains a challenging research problem as the performance gap between automatic speech recognition and human listeners is widest when speech is corrupted with noise. The work presented in this section is focussed on methods that model the uncertainty in the observed or reconstructed (cleaned) speech features when the clean speech signal is corrupted with noise from an unknown source. In addition to the uncertainty-based methods presented here, we have continued the work on noise robust feature extraction using weighted linear prediction [1].

Missing feature approaches

The so called missing-feature methods are a special case of methods that use observation uncertainty or reliability in order to improve speech recognition performance in noisy conditions. The methods, which draw inspiration from the human auditory system, are based on the assumption that speech corrupted by noise can be divided to speech-dominated i.e. reliable regions and noise-dominated i.e. unreliable regions as illustrated in Figure 8.3. The clean speech information corresponding to the unreliable regions is assumed missing,

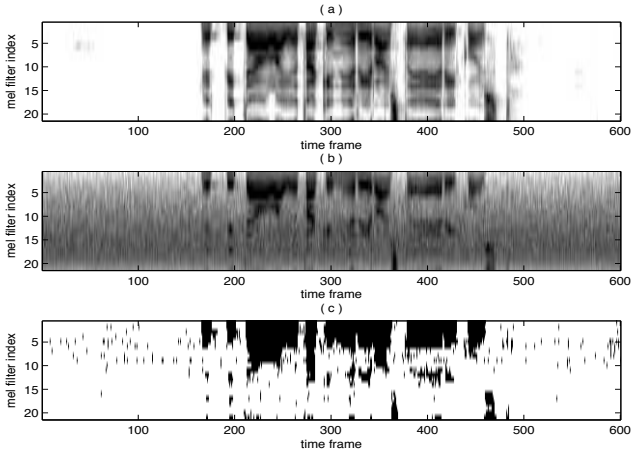

Figure 8.3: Logarithmic mel spectrogram of (a) an utterance recorded in quiet environment and (b) the same utterance corrupted with additive noise. The noise mask (c) constructed for the noisy speech signal indicates the speech dominated regions in black and the noise dominated regions in white.

which means that under additive noise assumption, the observed values determine an upper bound for the unobserved clean speech features but contain no further information regarding the missing values. In noise-robust speech recognition, the missing clean speech information is either marginalised over or reconstructed using missing-feature imputation techniques [2]. The reconstruction approach was compared with other noise-robust speech recognition methods in [3].

Reconstruction methods are based on modelling the statistical dependencies between clean speech features and using the model and the reliable observations to calculate clean speech estimates for the missing values. Recent improvements to missing-feature imputation are

due to modelling the temporal dependencies between clean speech features in consecutive frames. Processing the noisy speech in windows that span several time frames was first proposed in the exemplar-based sparse imputation (SI) framework [4]. SI outperformed the conventional GMM-based imputation method that used frame-based processing. Window-based processing was later introduced in the GMM-based framework in [5], and to investigate other approaches to temporal modelling, a nonlinear state-space model (NSSM) based framework was developed for missing-feature reconstruction in [6]. Both the window-based GMM and the NSSM imputation method outperformed frame-based GMM imputation in all experiments and outperformed SI when evaluated under loud impulsive noise.

In addition to work on improving the core missing feature methods, we have studied missing feature methods in models of human hearing. Related to this work, we proposed a model that explains the speech recognition performance of human listeners in a binaural listening scenario [7]. Furthermore, we have applied the missing-feature reconstruction methods developed for noise-robust speech recognition to extending the bandwidth of narrowband telephone speech to the high frequency band [8] and the low frequency band [9]. The latter study won the International Speech Communication Association award for the best student paper in Interspeech 2011.

Modelling uncertainty in reconstruction

In addition to using reliability estimates to determine reliable and unreliable features in missing-feature reconstruction, we have studied using another type of reliability estimates to improve the speech recognition performance when reconstructed or otherwise enhanced speech data is used. First, we have studied uncertainty estimation in the context of sparse imputation [10, 11]. Unlike the parametric methods that model clean speech using a GMM or NSSM, for example, the exemplar-based sparse imputation method does not provide for calculating a full posterior for the reconstructed features. We therefore investigated using a number of heuristic measures to represent the uncertainty related to the SI reconstruction performance. Similarly, we have developed a number of heuristic uncertainty measures for the exemplar-based sparse separation technique that uses a speech and noise dictionary to estimate clean speech features based on the noisy observations [12].

References

[1] S. Keronen, J. Pohjalainen, P. Alku, and M. Kurimo, Noise robust LVCSR feature extraction based on extended weighted linear prediction. Proc. INTERSPEECH, 2011.

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

[3] S. Keronen, U. Remes, K. J. Palomäki, T. Virtanen and M. Kurimo, Comparison of noise robust methods in large vocabulary speech recognition, Eusipco 2010.

[4] J. F. Gemmeke, B. Cranen, and U. Remes (2011). Sparse imputation for large vocabulary noise robust ASR. Computer Speech and Language, vol 25, issue 2, pp. 462-479, 2011.

[5] U. Remes, Y. Nankaku, and K. Tokuda, GMM-based missing feature reconstruction on multi-frame windows. Proc. INTERSPEECH, pp. 1665-1668, Florence, Italy, August 2011.

[6] U. Remes, K. J. Palomäki, T. Raiko, A. Honkela and M. Kurimo, Missing-feature reconstruction with bounded nonlinear state-space model, IEEE Signal processing letters, 18(10), 563-566, 2011

[7] K. J. Palomäki and G. J. Brown A computational model of binaural speech recognition: role of across-frequency vs. within-frequency processing and internal noise, Speech Communication, 53(6), 924-940, 2011

[8] H. Pulakka, U. Remes, K. J. Palomäki, M. Kurimo, P. Alku, Speech bandwidth extension using Gaussian Mixture Model-based estimation of the highband Mel spectrum. Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP 11), Prague, Czech Republic, May 22-27, 2011.

[9] H. Pulakka, U. Remes, S. Yrttiaho, K. Palomäki, M. Kurimo and P. Alku, Low-frequency bandwidth extension of telephone speech using sinusoidal synthesis and gaussian mixture model. In Proc. Interspeech 2011, Florence, Italy, Aug. 28-31, 2011.

[10] J. Gemmeke, U. Remes and K. J. Palomäki, Observation uncertainty measures for sparse imputation, Interspeech 2010.

[11] H. Kallasjoki, S. Keronen, G. J. Brown, J. F. Gemmeke , U. Remes and K. J. Palomäki, Mask estimation and sparse imputation for missing data speech recognition in multisource reverberant environments, in International Workshop on Machine Listening in Multisource Environments, 2011.

[12] H. Kallasjoki, U. Remes, J. F. Gemmeke, T. Virtanen and K. J. Palomäki, Uncertainty measures for improving exemplar-based source separation, in Proc. Interspeech 2011.

8.4 Constraining and adapting language models

Early speech recognition systems used rigid grammars to describe the recognized language. Typically the grammar included a limited set of sentences used to command the system. Such language models do not scale for large vocabulary continuous speech recognition. Therefore modern recognizers, including the Aalto University recognizer, use statistical language models.

Constrained command languages are still useful in some spoken dialog applications, where commands are important to be recognized correctly, especially if the system cannot be adapted to a specific user group. We have succesfully built statistical language models from command grammars, modeled in Backus-Naur Form (BNF). Language models built in this way enable fast decoding and near perfect recognition accuracy.

When large-vocabulary speech recognition is applied in a specialized domain, the vocabulary and speaking style may substantially differ from those in the corpra that are available for Finnish language. Using additional text material from the specific domain, when estimating the language model, is beneficial, or even necessary for proper recognition accuracy. We have applied speech recognition to medical transcription. A huge collection of dental reports was received from In Net Oy, for estimating a language model specific to dental dictation. User tests are underway, but our benchmarks indicate large differences in accuracy between different users.

Collecting domain-specific texts is time-consuming and usually there's not enough data available to estimate a reliable language model. Most of the times we have to use the little in-domain data we have to adapt the general language model.

In a project aimed at developing a mobile dictation service for lawyers, we used law-related texts to train an in-domain language model [1]. Adapting the general language model with the in-domain model usually gave better results than just using either model separately. One of the key challenges of the project was still to find proper adaptation data. Even though the adaptation texts are of the targeted domain, the language of the real-life dictations can still be significantly different than the written text.

Language model adaptation usually consists of mixing or combining the probabilities of the general language model with the in-domain model. The most simple and popular LM adaptation method is linear interpolation. Linear interpolation is performed by simply calculating a weighted sum of the two models probabilities.

We have experimented with a more sophisticated LM adaptation method, which uses the information theory principle of maximum entropy (ME) to adapt the general language model with the in-domain model [2]. The key to this approach is that the global and domain-specific parameters are learned jointly. Domain-specific parameters are largely determined by global data, unless there is good domain- specific evidence that they should be different. We tested the method on English and Estonian broadcast news and experiments showed that the method consistently outperformed linear interpolation. The main drawback with this method is that it's very memory and time consuming.

The implementation of ME language model adaptation is freely available as an extension to the SRI language modeling toolkit [3].

References

[1] A. Mansikkaniemi. Acoustic Model and Language Model Adaptation for a Mobile Dictation Service. *Master's thesis, Aalto University*, 2010.

[2] T. Alumäe and M. Kurimo, Domain Adaptation of Maximum Entropy Language Models, Proceedings of the ACL 2010, Uppsala, Sweden, July 2010.

[3] T. Alumäe and M. Kurimo, Efficient Estimation of Maximum Entropy Language Models with N-gram features: an SRILM extension, Proceedings of Interspeech 2010, Chiba, Japan, September 2010.

8.5 Speech retrieval and indexing

Speech retrieval techniques enable users to find segments of interest from large collections of audio or video material. Automatic speech recognition is used to transforms the spoken segments in the audio to textual form. Information retrieval (IR) methods are used to index the text, and to perform searches on the material based on query words typed by the user. Since the amount of information in spoken form is very large and ever increasing, the methods developed have to be fast and robust to be able to process large amounts of variable quality material.

One complication in the speech retrieval process is the fact that the speech recognizer output will always have erroneous words. A special problem for speech retrieval are *out-of-vocabulary* (OOV) words – words that are not in the list of words the speech recognizer knows. Any OOV word in speech can not be recognized, and is replaced by similarly sounding but usually unrelated word. Since query words are chosen to be discriminative, they are often rare words such as proper names. But rare words are often also OOV, since the recognizer vocabulary is chosen so that a number of most common words are included.

This problem can be solved by using recognition units that are smaller than words, but that are large enough to be able to model the language. Morphs produced by the Morfessor algorithm have been proven to work well as such units. The speech recognizer language model is trained on a text corpus where the words are split to morphs, and the recognizer is then able to transcribe any word in speech by recognizing its component morphs. It is possible to join the morphs to words and use traditional morphological analyzers to find the base forms of the words for indexing. But since there will still be an amount of errors in the morph transcripts, especially when the spoken word is previously "unseen", a word that did not appear in the language model training corpus, using morphs as index terms will allow utilizing the partially correct words as well. In this case, query words are also split to morphs with Morfessor. Experiments using Finnish radio material show that morphs and base forms work about equally well as index terms, but combining the two approaches gives better results that either alone [1]. Table 8.1 shows an example how OOV words are recognized with word and morph language models.

Table 8.1: Example recognition results of two unseen query words at two different locations each. With the morph language model, it is possible to recognize correctly at least some of the morphs, which will match morphs in the query. With the word language model, the words are replaced by unrelated words.

Query word	Iliescun		Namibian	
- Translation	*Iliescu's*		*Namibia's*	
Morph query	ili escu n		na mi bi an	
Morph LM rec.	n ilja escu	ili a s kun	ami bi an	na min pi an
Word query	iliescun		namibia	
Word LM rec.	lieskoja	eli eskon	anjan	namin pian
Word lemmas	lieska	eli elää esko	anja	nami pian pia
- Translation	*flame*	*or live Esko*	*Anja*	*candy soon Pia*

Audio and video is typically distributed as a flow of material without any structure or indicators where the story changes. Thus, before indexing, the material needs to be automatically segmented into topically coherent speech documents. This can be done e.g. by measuring the lexical similarity of adjacent windows. Morphs were found to help in

the segmentation task as well when processing ASR transcripts [1].

Retrieval performance can be further improved by utilizing alternative recognition candidates from the recognizer [1]. 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.

References

[1] V.T. Turunen, and M. Kurimo, Speech retrieval from unsegmented Finnish audio using statistical morpheme-like units for segmentation, recognition, and retrieval. *ACM Transactions on Speech and Language Processing*, Vol. 8, No. 1, pp. 1–25, October 2011.

[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.