Chapter 9

Speech recognition

Mikko Kurimo, Panu Somervuo, Kalle Palomäki, Vesa Siivola, Teemu Hirsimäki, Janne Pylkkönen, Simo Broman, Ville Turunen, Sami Virpioja

9.1 The speech recognition tasks and systems

This chapter is divided into four categories that describe our research activities in: **1.Acoustic modeling, 2.Language modeling, 3.Large vocabulary decoders, and 4.Speech retrieval**. The division is natural both because it covers four of the major subfields in speech recognition research and because it describes the main components of a typical large vocabulary continuous speech recognition (LVCSR) system (Figure 9.1). The acoustic models produce the probabilities of different phonemes, the language models take into account the co-occurrence probabilities of different words or morphemes, the decoder joins these two streams of information into recognition hypothesis, and the retrieval engine utilizes these outputs to represent the speech in a convenient form for searching and browsing. Thus, all our research topics focus on the same framework and can be integrated into a single working LVCSR system.

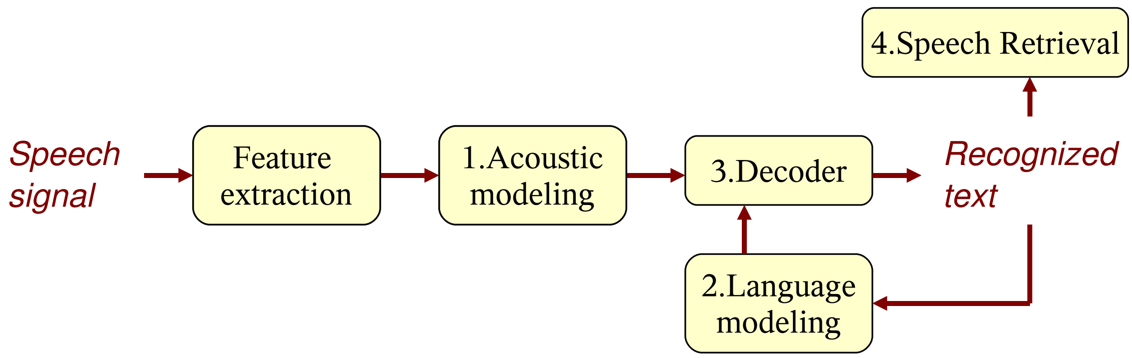

Figure 9.1: The main components of the LVCSR system.

Our goal in LVCSR research has for several years been to develop new machine learning algorithms for each of the subfields and build a complete state-of-art recognizer to evaluate the 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 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, 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 LVCSR and language modeling in different languages. This challenge project is funded by EU's PASCAL network.

9.2 Acoustic modeling

Phoneme modeling and speaker adaptation

Acoustic modeling in automatic speech recognition (ASR) 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.

After the feature extraction the feature sequence 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 9.2. 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. Even though the Gaussian mixture components are restricted to have only diagonal covariance matrices, the number of parameters with such a complex acoustic model in a typical state-of-the-art ASR system is very high, in order of millions of parameters. This gives emphasis to proper complexity control, so that we get the most out of the available training data.

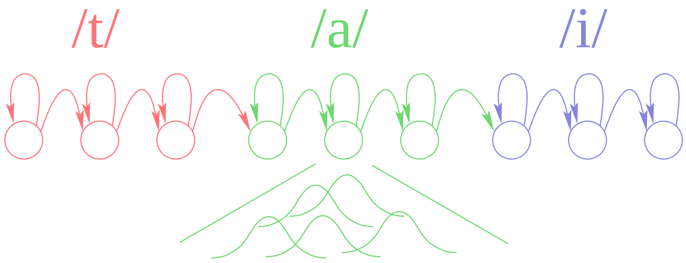

Figure 9.2: Each phoneme is modeled with a hidden Markov model, usually consisting of three states. The state distributions are modeled by Gaussian mixture models.

The problem of selecting the optimal model complexity is a difficult one, but it can be avoided by using some other model training criterion instead of the usual maximum likelihood (ML) principle. In [1] two more advanced training principles, maximum a posteriori (MAP) and variational Bayesian (VB), were compared against the ML principle. These two methods can avoid overfitting in case of too complex a model, which is the major drawback in ML training. This was also validated experimentally, where the speech recognition performance started to degrade with ML trained models when the number of parameters was increased, whereas MAP and VB trained models continued to work well. The VB principle can also be used to select the proper model complexity in respect to the training data, without using auxiliary data.

Hidden Markov models have several drawbacks with respect to speech modeling. One of those is the modeling of the durations of speech segments. Standard HMMs allow only minimal modeling of duration variations, although in some languages (e.g. in Finnish) the durations can be the main cues in discriminating between certain phonemes. To better take the durations into account we experimented in [2] several extensions to standard HMMs which allow more precise models for the segmental durations. It was found out that already a relatively simple duration model was enough to improve the speech recognition results.

In the past most of our speech recognition experiments have been carried out with

Method	Word error rate (%)	Phoneme error rate (%)
Baseline	30.5	9.8
VTLN	29.3	9.1
cMLLR	25.3	7.3
SAT/cMLLR	24.2	6.8

Table 9.1: Speech recognition results for several adaptation methods: No adaptation (Baseline), Vocal Tract Length Normalization (VTLN), Constrained Maximum Likelihood Linear Regression (cMLLR) and Speaker Adaptive Training with cMLLR (SAT/cMLLR).

speaker dependent models, meaning the acoustic models have been trained specifically for one person. Recently we have been able to move to more demanding speaker independent experiments, which is also more realistic in view of many applications. The lack of speaker dependency adds further demands for the acoustic models. We have therefore implemented several adaptation techniques to our speech recognizer, and tested their effectivity with our speech data. Some results are shown in Table 9.1.

Recognition of reverberant speech

In the acoustic modeling for large vocabulary continuous speech recognition mostly speech in relatively noise free condition was concentrated (see Sect. 9.2). In the field of noise robust speech recognition, we have been developing techniques to recognition of reverberant speech jointly with the University of Sheffield [4]. Our technique is based to missing data approach [5], in which a conventional Gaussian mixture model classifier is adapted to allow different treatments of reliable and unreliable regions of speech. In our approach the regions of speech spectrum, which are either relatively clean or badly contaminated by reverberation are indexed and used to construct a time frequency mask to the missing data speech recognizer. Masks are produced by applying modulation filtering to detect strong speech regions not contaminated by reverberation (see Fig. 9.3). Furthermore, we were able to improve the performance slightly by combining the missing data recognizer to a conventional recognizer using cepstral features. More information about techniques to handle reverberation in the auditory scene analysis as well as in speech recognition can be read from our recent review article [3].

References

[1] P. Somervuo: Comparison of ML, MAP, and VB based acoustic models in large vocabulary speech recognition In *Proceedings of the 8th International Conference on Spoken Language Processing* (Interspeech 2004), October 4–8, 2004, Jehu Island, Korea, pp. 701–704.

[2] J. Pylkkönen and M. Kurimo: Duration Modeling Techniques for Continuous Speech Recognition In *Proceedings of the 8th International Conference on Spoken Language Processing* (Interspeech 2004), October 4–8, 2004, Jeju Island, Korea, pp. 385–388.

[3] 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, to appear in June 2006.

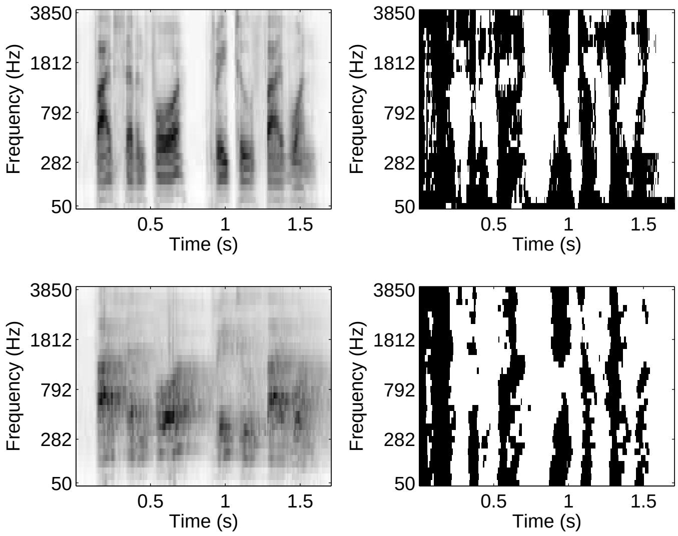

Figure 9.3: Auditory spectrograms (left panels) and missing data masks (right panels). The spectrogram for a "clean" unreverberated (top-left) and a highly reverberant (bottom left) speech utterance are shown. The right panels show the corresponding missing data masks for the reverberant utterance. Firstly, an "oracle mask" based on prior knowledge (top-right) of the reverberated regions shows how an (nearly) ideal mask should look like. Secondly, a mask produced using our model (with no prior knowledge) is shown. Black and white regions indicate reliable and unreliable regions, respectively.

[4] K. J. Palomäki, G. J. Brown and J. Barker, Recognition of reverberant speech using full cepstral features and spectral missing data Accepted for publication in *Proc. ICASSP 2006.*

[5] 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.

9.3 Language modeling

Splitting words into fragments

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 recognition tasks indicate that an unsupervised data-driven splitting algorithm called Morfessor (see Section 10.1) can produce word fragments that work even better in speech recognition than morphemes based on Finnish grammar [1, 2].

Since the Morfessor algorithm is language independent, it can also be applied to speech recognition of other languages. Experiments in Turkish and Estonian recognition tasks confirm the result that models based the Morfessor algorithm improve recognition accuracy.

A growing method for constructing an n-gram model

The length of the word history used by the n-gram model is traditionally set to a fixed n. For $n > 3$ this often leads to prohibitively big models. We have developed an algorithm based on the Minimum Description Length principle [3], which learns a suitable word history length for each case [4]. The factors affecting the choice of histories are: 1) Does the model get much better if we use a longer word history for modeling an n-gram? and 2) Do we have enough data to estimate the probabilities for the longer history? This method can make considerably smaller n-gram models which equal modeling power of the fixed n models.

A related method is the pruning of n-gram models, for example entropy based pruning [5]. The benefits of our approach compared to pruning methods are that at no time we need to store the full model. This allows us to train very high order models. Our experiments show, that the growing method seems to outperform the entropy based pruning in practically all experiments [6]. For example in Finnish speech recognition experiments, the growing method gives at least 15% lower word error rate (relative) for reasonable n-gram model sizes, when both methods use equal model size. The n-gram models can be efficiently stored in a tree structure (Fig. 9.4).

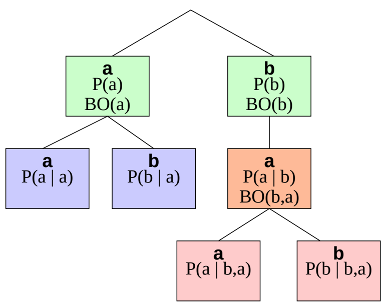

Figure 9.4: The tree structure for storing an n-gram model.

Combining the growing method and clustering

In addition to the pruning of the n-gram models, a common way to decrease the size of the n-gram models is clustering of model units or sequences of them. In a similar manner that the MDL principle can be used to choose a suitable length of n-gram history for each case, it can be used to insert n-gram histories that give similar predictions into same equivalence classes [7]. Compared to the baseline of the growing method, for a model of an equal size, some accuracy may be lost, but substantially more n-grams can be included into the model.

In order to make the clustering computationally fast, the number of different model units cannot be very large. Suitable small lexicons are easy to construct for any language with the Morfessor algorithm (Section 10.1). Preliminary experiments show that with some optimizations, even extensive searches for the nearest history clusters are possible. This differs from e.g. one related method [8], where only nearby parts of the tree structure are searched for similar prediction distributions.

Combining methods for language models

In many task the best language modeling results have been achieved when different language models have been used together [9]. Several combination methods have been presented in the literature, but a thorough investigation of the methods has not been done.

In [10, 11], the combination methods that have been used with language models are studied. Also, a new approach based on likelihood density function estimation using histograms is presented. In addition to theoretical consideration, four combining methods for four language models are evaluated in speech recognition experiments and word prediction experiments using Finnish news articles.

In the perplexity experiments, all combining methods produced statistically significant improvement compared to the 4-gram model that worked as a baseline. The best result, 46 % improvement to the 4-gram model, was achieved when combining several language models together by using the new bin estimation method. In the speech recognition experiments, 4 % reduction to the word error and 7 % reduction to the phoneme error was achieved.

References

[1] T. Hirsimäki, M. Creutz, V. Siivola, M. Kurimo, S. Virpioja, J. Pylkkönen, Unlimited vocabulary speech recognition with morph language models applied to Finnish. *Computer Speech & Language*, accepted for publication in 2005.

[2] T. Hirsimäki, M. Creutz, V. Siivola and M. Kurimo: Morphologically Motivated Language Models in Speech Recognition, In *Proceedings of the International and Interdisciplinary Conference on Adaptive Knowledge Representation and Reasoning* (AKRR05), June 15-17, 2005, Espoo, Finland, pp. 121-126.

[3] Rissanen, J., 1994. *Language Computation.* American Mathematical Society, Ch. Language Acquisition in the MDL Framework.

[4] V. Siivola, "Building compact language models incrementally," in *Proceedings of Second Baltic Conference on Human Language Technologies*, 2005, pp. 183–188.

[5] Stolcke, A., 1998. Entropy-based pruning of backoff language models. *In Proc. DARPA Broadcast News Transcription and Understanding Workshop.* pp. 270–274.

[6] Siivola, V., Pellom, B., 2005. Growing an n-gram model. *In Proc. Interspeech 2005.* pp. 1309–1312

[7] Virpioja, S., 2005. New methods for statistical natural language modeling. Master's thesis, Department of Computer Science and Engineering, Helsinki University of Technology.

[8] Siu, M., Ostendorf, M., 2000. Variable n-grams and extensions for conversational speech language modeling. *IEEE Transactions on Speech and Audio Processing*, 8(1):63–75.

[9] Kurimo, M., Zhou, B., Huang, R., Hansen, J.H.L., 2004. Language modeling structures in audio transcription for retrieval of historical speeches. *In Proc. EUSIPCO 2004.*

[10] Broman, S., 2005. Combining methods for language models in speech recognition. Master's thesis, Helsinki University of Technology.

[11] Broman, S., Kurimo, M., 2005. Methods for combining language models in speech recognition. *In Proc. Interspeech 2005.* pp. 1317–1320

9.4 Large vocabulary decoder

The goal of the speech recognition is to find the word sequence that is the most probable one given the acoustic model, language model and the observed speech. Because the number of possible word sequences is extremely high, the search is performed incrementally in time, and improbable hypotheses are pruned as early as possible. The module responsible for performing this search is called the decoder.

During the recent years, we have actively developed a decoder for very large vocabularies. In order to take the acoustic dependencies better into account, the stack based decoder has been replaced by an efficient time-synchronous token-pass decoder [1]. The efficiency is derived from a compact search network which can utilize the redundancies in the acoustic models. The decoder is able to model correctly also the context dependent phonemes which occur across lexical units. Compared to our previous decoder, this results in a 24% relative improvement of the phoneme error rate.

The efficiency of decoders continues to be an important issue in speech recognition, as more and more complex models of acoustics and language are used to obtain the possible best recognition accuracy. Some new ways to restrict the search space without affecting the recognition accuracy too much were developed in [2]. These so called pruning criteria use different information available during the search to discard those path hypotheses which no longer seem feasible. The research also resulted in a method with which we can avoid hand tuning the numerous parameters affecting the efficiency/accuracy tradeoff in the decoding process.

References

[1] J. Pylkkönen: An Efficient One-pass Decoder for Finnish Large Vocabulary Continuous Speech Recognition, In *Proceedings of the 2nd Baltic Conference on Human Language Technologies* (HLT'2005), April 4–5, 2005, Tallinn, Estonia, pp. 167–172.

[2] J. Pylkkönen: New Pruning Criteria for Efficient Decoding, In *Proceedings of the 9th European Conference on Speech Communication and Technology* (Interspeech 2005), September 4–8, 2005, Lisboa, Portugal, pp. 581–584.

9.5 Spoken document retrieval

Speech retrieval and indexing

One important application of automatic speech recognition is spoken document retrieval which means the task of finding interesting segments from recorded speech. Large amount of information is produced in spoken form, for example radio and TV broadcasts, and there is a need for tools that can be used to search this data. Spoken document retrieval systems combine speech recognition and information retrieval technologies. The special properties of the Finnish language, such as the large number of inflected word forms, affect both of these parts and methods developed for other languages cannot be used as such. Our research is focused on retrieval of Finnish speech, but the methods are hoped to work also on other languages with similar properties.

Previously, word-based and phone-based approaches have been used. The former suffers from limited vocabulary and the latter from high error rates. In [1, 2], we presented a baseline retrieval system for Finnish that uses the speech recognizer based on morpheme-like subword units. The recognizer can achieve low error rates while providing a potentially unlimited vocabulary. Retrieval performance of spoken news was found to be close to that of the human reference transcripts. The morpheme like units were found to work well also as index terms, providing equal performance to base formed words.

Recognition errors degrade retrieval performance, but there are measures that can be used to reduce their effect. For example, the recognizer can be modified to include alternative recognition results in the transcripts, or queries can be expanded by adding relevant words from a parallel text corpus. Query expansion was found to bring the level of performance to the same as text document retrieval, even for transcriptions with relatively high error rates. [3, 4]

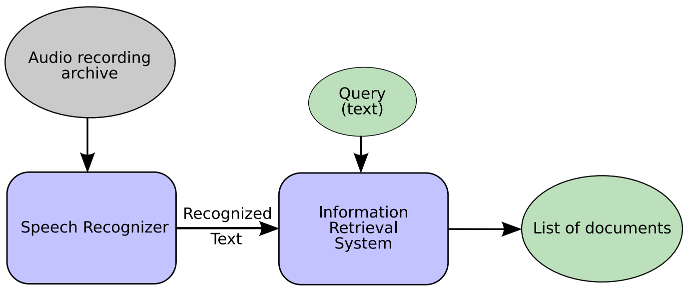

Figure 9.5: Overview of a spoken document retrieval system.

Speech segmentation

The development of automatic segmentation methods of speech and audio is increasingly important to allow automatic handling of growing archives of spoken audio, e.g. recorded meetings, radio or television programs. Audio material can be segmented based on various levels of description. On metadata level audio can be classified e.g. to speech vs. multiple classes of non-speech. Furthermore, segments containing only speech can be classified based on gender, speaker identity and, finally, into subunits of speech, such as, sentences, words or phonemes. Segmentation can be performed either in a supervised or unsupervised manner. In the supervised segmentation, the task is to align temporal structure of speech to the existing transcription. In the unsupervised segmentation the transcript does not exist, and the recognizer classifies the segments freely.

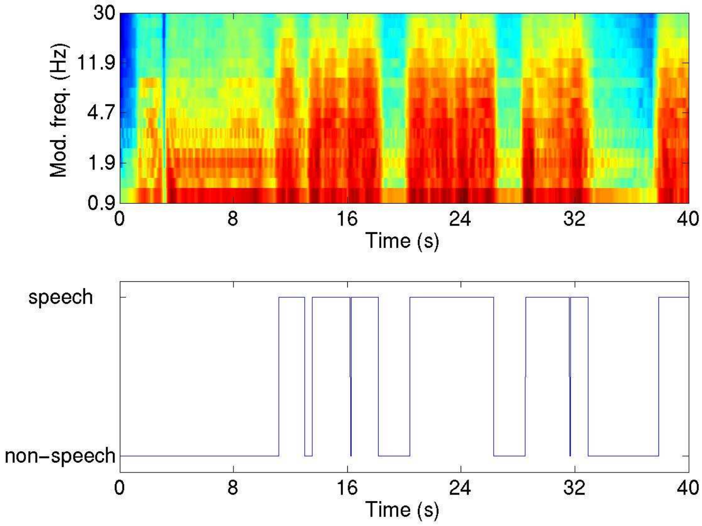

Figure 9.6: Speech modulation spectrum (top) against target classification (bottom).

The group's speech recognition tools [1] were applied to supervised phoneme level segmentation to align existing transcriptions to the corresponding speech audio. This line of research was extended in a student project to speech that is only partially transcribed. Moreover, practical speech recognition tasks have prompted researchers in the group to develop unsupervised techniques [5] to speaker segmentation. These methods were also used to address the needs of other speech researchers in the Helsinki University, Tampere University of Technology and University of Turku.

A new research project in speech segmentation in the metadata level was initiated during April 2005. In this project, we are developing techniques to extract speech segments from audio stream, and speech segments based on gender in an unsupervised manner. The project was started with development of feature techniques using a common Gaussian mixture model classifier. In our new approach, we have applied two types of feature presentations of speech, first, which depicts short-term ($\approx$ 16 ms) speech spectrum and the second that describes long term temporal modulations ($\approx$ 1 s) of speech applying a computation of modulation spectrum (see Fig. 9.6).

References

[1] M. Kurimo, V. Turunen and I. Ekman: An Evaluation of a Spoken Document Retrieval Baseline System in Finnish, *ICSLP*, 2004.

[2] M. Kurimo, V. Turunen and I. Ekman: Speech Transcription and Spoken Document Retrieval in Finnish, *Machine Learning for Multimodal Interaction* Revised Selected Papers of the MLMI 2004 workshop. Lecture Notes in Computer Science, Vol. 3361, pages 253–262, 2005

[3] M. Kurimo and V. Turunen: To recover from speech recognition errors in spoken document retrieval, *Interspeech 2005*, pages 605–608, 2005.

[4] V. Turunen, Spoken document retrieval in Finnish based on morpheme-like subword units, M.S. thesis, Helsinki University of Technology, Espoo, Finland, 2005.

[5] L. Wilcox, F. Chen, D. Kimber and V. Balsubramaman Segmentation of speech using speaker identification. *Proc. ICASSP 1994*, pp. I-161-I-164.