1 Acoustic modeling

The general goal of automatic speech recognition (ASR) is to understand normal human speech and then to be able to perform some task based on this understanding. One application of ASR is a dictation system which converts spoken sentences into their written forms. In this sense the speech recognition can be defined as the mapping from the continuous acoustic signal to the discrete set of symbols.

There are several reasons for the difficulties in speech recognition. Natural speech has variations in many levels. In addition to that different speakers have different voices, there is also considerable variation in the voice of a single speaker. In normal conversation, some parts of the words may be emphasized more than others depending on the context. The loudness and the pitch of the voice may change and also the speaking rate may vary. Even if the speaker tries to use as steady a voice as possible, no two uttered sounds are generally equal. Therefore, in the speech recognition system, some limitations are usually made concerning the nature of the speech to be recognized. These limitations may include the number of the speakers, the size of the vocabulary, the amount of the noise in the speech, and the assumption that the input will always be speech.

Our projects in automatic speech recognition are aimed both to use the recognition system as a test bench for the neural network algorithms developed in the laboratory and to develop the system itself as a pilot application of the neural networks. Besides developing new **recognition algorithms**, we have also investigated new **acoustic features** and **context modeling**. Examples of the applications where we have used SOM and LVQ algorithms are shown in Table 1 and Figure 1.

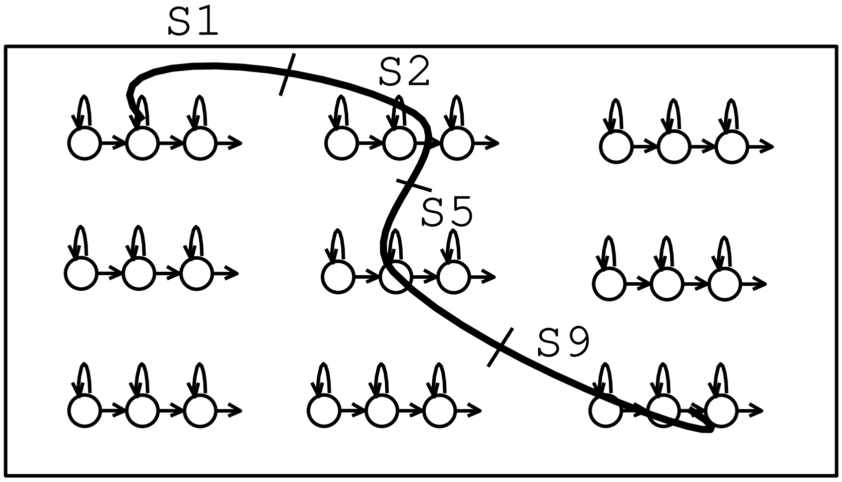

Figure 1: Competitive-learning segment models on the SOM. Each map node is associated with an HMM (with three states in this case) instead of a traditionally used single feature vector. The thick line represents the Viterbi segmentation of one input sequence. This corresponds to the best matching unit (BMU) search. The models of the BMUs and neighboring units are then updated by the corresponding segments.

The block diagram of a speech recognition system is shown in Figure 2. The recognition is based on connecting the hidden Markov models (HMMs) of the phonemes to decode the phoneme sequences of the spoken utterances. The output density function of each state in each model is a mixture of multivariate Gaussian densities. We have used the following scheme for the training of the models [1]. SOM is used first for initializing the phoneme-wise codebooks. Each model vector becomes then a mean vector of a

Gaussian kernel. After initialization, the training is continued by segmental-SOM or K-means algorithm. Segmental-LVQ is then applied for error-corrective training in order to obtain better phoneme discrimination. In the context of mixture density HMMs, we have developed methods for speeding up the recognition [1,3] based on the SOM structure.

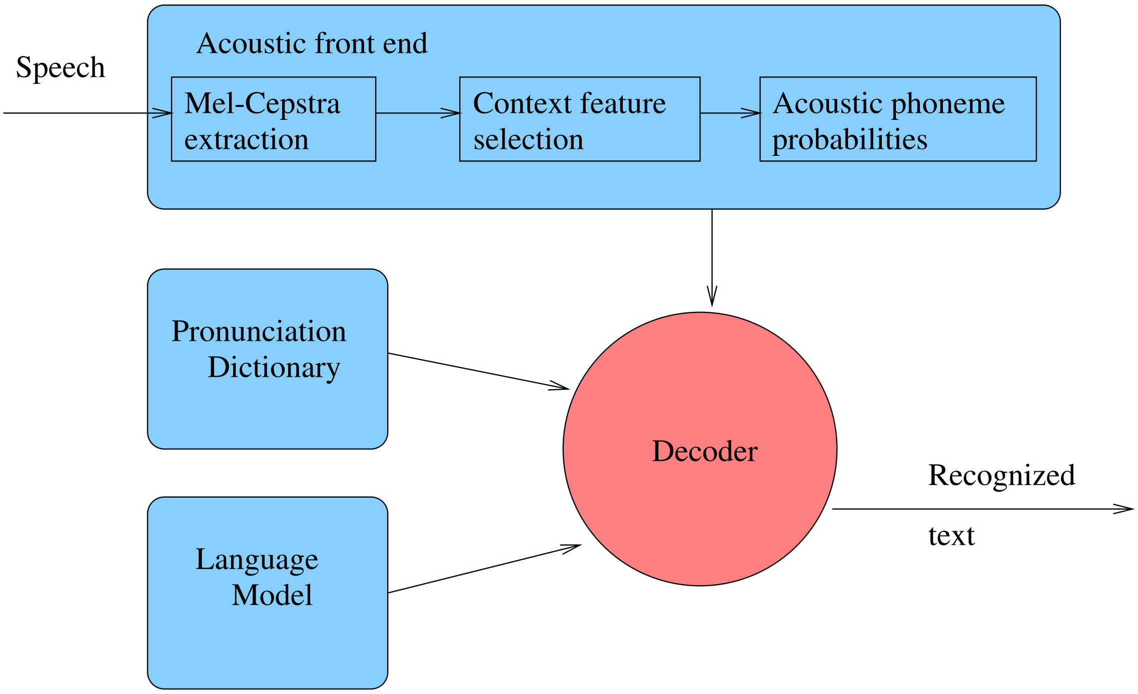

Figure 2: Overview of an ASR system.

Model associated with the SOM node:	Application:
1. feature vector	kernel means of mixture densities [1]
2. feature vector sequence	variable-length word templates [3]
3. hidden Markov model	set of (non-linguistic) speech segments [3]
4. symbol string	learning pronunciation dictionary [3]
5. word n-gram	word cluster in a language model [2]
6. word histogram	language model of a topic cluster [2]

Table 1: ASR-related SOM and LVQ applications.

As a promising future alternative for acoustic models in speech recognition an active research topic in the laboratory has also been the development of continuous state-space models of speech and Bayesian ensemble learning for latent variables. For more information on these topics see the chapter corresponding to the Bayesian modeling group.

Besides the recognition of speech, we have used our models also for the segmentation of new large Finnish speech corpora. The segmentation of the new data is an essential first step before the new material can be used for training. This work is related to the national USIX research program and has been helpful for the other participating Finnish speech research groups who are working on the same database.

2 Language modeling

The output of the phonemic vocabulary-free recognizer will inevitably contain some errors. Our current research is focused towards **large vocabulary continuous speech recognition** (LVCSR) systems and **language modeling**. The role of the language model is to control the search of the best phoneme or word sequence and improve the recognition. Since the best modeling methods are language specific, we cannot simply use the same models which have given good results e.g. for English. We have to cope with the special characteristics of the Finnish language which include e.g. a relatively free word order and a very large recognition vocabulary due to the number of inflected word forms and compound words. Some these problems are common to other non-english languages as well.

In order to better test the new methods and algorithms we are currently developing a new efficient decoder for the LVCSR task. This will be integrated in our speech recognition system.

One research topic has been how to better estimate the parameters of a language model. Since the models can consist of tens of millions of parameters, the parameter estimation is very sensitive to training methods and peculiarities of the training data. By carefully compressing the language model down to much fewer parameters, we increase the model's robustness and its ability to generalize for unseen case. One way to reduce the parameter count is to cluster similar words to one cluster and operate on these clusters instead of individual words [2].

An emerging new research topic is the use of the efficient language processing tools developed in the laboratory (WEBSOM) to organize language models based on the topical structure of the discourse [2]. The objective is to increase the language modeling accuracy and to obtain improved speech recognition results by automatically detecting and focusing into the best available language model for the recognition task at hand. This work is done in a close collaboration with the Natural language modeling group (see Section **??**).

References

[1] M. Kurimo. Using Self-Organizing Maps and Learning Vector Quantization for Mixture Density Hidden Markov Models. *PhD Thesis*. Helsinki University of Technology, Neural Networks Research Centre, 1997.

[2] V. Siivola, M. Kurimo, and K. Lagus. Large vocabulary statistical language modeling for continuous speech recognition in Finnish. In *Proceedings of the 7th European Conference on Speech Communication and Technology*, volume 1, pages 737–730, 2001.

[3] P. Somervuo. Self-Organizing Maps for Signal and Symbol Sequences. *PhD Thesis*. Helsinki University of Technology, Neural Networks Research Centre, 2000.