USING MULTI–LEVEL SEGMENTATION COEFFICIENTS TO IMPROVE HMM SPEECH RECOGNITION

*Kai Hübener**

University of Hamburg
Computer Science Dept.
huebener@informatik.uni–hamburg.de

ABSTRACT

This paper presents a new kind of acoustic features for HMM speech recognition. These features try to capture phone–specific segmentation information using multiple temporal resolutions. Experiments show that word accuracy can be improved by 7% when combining these features with traditional mel–cepstral coefficients in a speaker–independent word recogniser. This improvement is mostly due to a reduced number of insertion and deletion errors.

1. INTRODUCTION

Extracting reliable acoustic features from speech signals is one of the prevailing issues in preprocessing for speech recognition. The impact of selecting robust features on overall recognition performance has only recently been demonstrated [9]. Most of today's systems use some variant of LPC–based or cepstral coefficients, thus assuming quasistationarity of the speech signal for the duration covered by a single feature vector.

Temporal changes in the spectra of speech signals are believed to play an important role in human perception. One way to capture this information is to use delta coefficients or differenced coefficients that measure the change in coefficients over time [3, 6]. Performance improvements have also been achieved by adding second differential coefficients [10]. Using multiple sets of delta coefficients obtained with different time delays led to further improvements [4]. Another way to incorporate temporal changes of speech spectra is the use of acoustic subword units as suggested in [1].

Temporal information is particularly suited for HMMs, since HMMs assume each frame to be independent of the previous frames, and dynamic features compensate for this assumption to some extent. However, the increased dimensionality may prevent a reduction of the recognition error rate, so additional should be robustly trainable. This does not necessarily apply for higher–order differential coefficients, as the results in [4] indicate.

Acoustic segmentation often provides valuable clues to improve recognition of diphtongs, nasals and stops. However, such information has proven difficult to integrate into the HMM framework since segmentation and classification are simultaneously optimized during maximum likelihood decoding. Attempts to separate these two steps usually result in degraded performance.

In the remainder of this paper, we present *segmental coefficients*, a new type of feature which effectively measures structural properties of speech. Experiments show that a 7% relative increase in recognition accuracy can be achieved for speaker–independent word recognition by combining mel–cepstral and segmental coefficients.

2. RELATED WORK

Explicit segmentation information has been used previously in HMM speech recognition [2, 8] to improve phone recognition by modifying inter–model transition probabilities or to segment the speech signal before using HMMs to determine the most likely phone sequence [7]. However, these approaches depend critically on the segmentation accuracy, since segmentation errors cannot be recovered later on. The approach described in this paper differs from previous approaches in four ways:

1. Segmental structure is considered an intrinsic feature of phones that is best modelled at the frame level.
2. Segmental structure is analysed using multiple temporal resolutions.
3. Standard training algorithms such as Viterbi or forward–backward training, can be used. Manual adjustment of weights is not necessary.
4. Segmentation information is used to improve word recognition instead of phone recognition.

3. SEGMENTAL COEFFICIENTS

Instead of separating segmentation and classification, we try to utilize the segmentation information in the acoustic signal

*The author is now with Siemens AG, ZT SN 5.

on the frame level in the form of additional coefficients during training as well as during the Viterbi alignment. This is different from previous approaches where segmentation information was only used during recognition and segmentation parameters had to be chosen manually.

To use segmentation information on the frame level, it has to be represented by a change function. Several such functions based on different multidimensional representations of the speech signal have been proposed to measure spectral variation. The one used here has also been successfully used for multilevel phone segmentation of speech in [5] when placing segment boundaries at local maxima of the change function.

We define the segmental coefficient $s_t(\sigma)$ at frame t as follows:

$$d_t^k(\sigma) = \sum_{\tau=1}^{N/2} c_{t+\tau}^k g'_{t+\tau}(\sigma) + c_{t-\tau}^k g'_{t-\tau}(\sigma) \quad (1)$$

$$s_t(\sigma) = \sqrt{\sum_{k=1}^{P} d_t^k(\sigma)^2} \quad (2)$$

where c_t^k is the k–th cepstral coefficient at frame t and P is the number of cepstral coefficients. The change function is smoothed by convolving it using the first derivative of a Gaussian filter of size N. Temporal resolution is controlled by σ. Basically, this function compares cepstral vectors separated by 2σ. Typical values for σ range between 2.0 ms and 20.0 ms. These values do not necessarily result in an optimal segmentation of the speech signal into phones as can be seen from Figure 1. Instead, they are selected to optimize recognition performance. Figure 1 shows a plot of five segmental coefficients at the top together with 12 mel–cepstral coefficients and log energy at the bottom. Vertical lines indicate phone boundaries generated by forced viterbi alignment.

Formant transitions as well as onsets and offsets will result in large values of $c_t(\sigma)$ since the distribution of spectral energy changes. Using a range of different values for σ allows capturing changes which occur on different time–scales, i. e. fast transitions like releases of stops as well as slow transitions between vowels and semivowels. To allow each time–scale to contribute equally to the output probability, each dimension was normalised to zero mean and unit variance.

4. EXPERIMENTS

Before finally integrating segmental coefficients into our speech recognition system, an experiment was made to use these coefficients for the segmentation of speech signals using a previously described search strategy [5]. The resulting segmentation lattices (see Figure 2) exhibits a strong correlation to a phonemic segmentation although a certain amount of errors is inevitable. The lattices can either be compared to manual phonemic segmentation to measure the accuracy of detected boundaries or used for restricting the Viterbi search in a subsequent HMM recogniser.

A test set of 704 utterances recorded from four female and seven male speakers was used. Using the data–driven path search algorithm described in [5], a segmentation accuracy of 90.4% was achieved for the best path in each lattice. This shows that segmental coefficients capture important information on the segmental structure of speech signals. The positions of segment boundaries were always within 10 ms away from the reference boundaries.

To assess the potential of segmental coefficients for speech recognition, a speaker–independent word recogniser was built using a commercially available HMM toolkit [10].

All experiments were carried out using 55 continuous, single–mixture, single–stream, context–independent HMMs each with three emitting states. Speech signals are blocked into 30 ms frames. Each frame is coded using 12 mel–cepstral coefficients, normalised log energy and optionally, the first differentials of these parameters. Training and testing utterances were taken from the Verbmobil corpus, a large collection of human–to–human spontaneous german dialogues. We use 643 utterances (75 minutes) for training and 109 utterances (9 minutes) for testing. The training utterances were spoken by 35 speakers (31m/4f). The test utterances were spoken by six different speakers (5m/1f). A bigram language model with a testset perplexity of 90 was used in all experiments.

All HMMs are initialised on phone labels generated by forced alignment using multiple pronounciations per word. Four iterations of forward–backward training are then used to train the models.

First, two reference systems were built using standard mel–cepstral coefficients (R1) and additionally delta cepstral coefficients (R2). The results obtained with these systems are given in Table 1.

System	# Coeffs	Acc	Ins	Del	Sub
R1	13	31.0	1.8	31.9	35.3
R2	26	40.6	5.6	18.9	34.9
S1	16	38.8	2.5	23.5	35.2
S2	27	43.5	4.4	18.8	33.3

Table 1: Recognition results

During HMM training it was noted that convergence is faster when using segmental coefficients. This can be explained by a better frame/state alignment when using segmentation information during the Viterbi alignment.

Next, segmental coefficients were used together with mel–cepstral coefficients (S1). As can be seen from Figure 3 recognition accuracy increases with the number of segmental coefficients used. This shows that no single temporal reso-

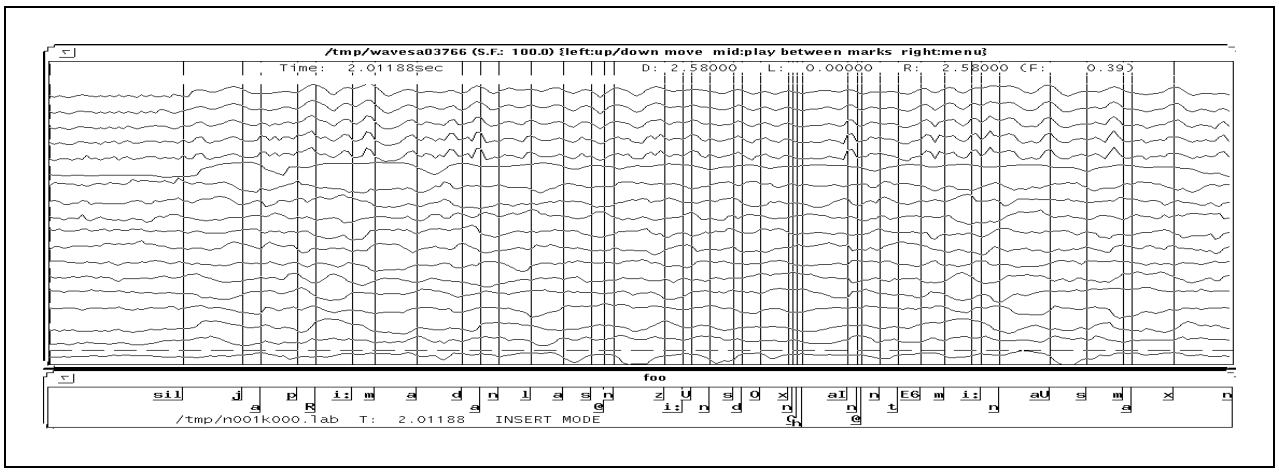

Figure 1: Coefficients and phone boundaries for the sentence N001K000

lution contains all the necessary information. When temproal resolution gets too coarse, performance drops again since phone–specific information starts to get distributed to the neighboring phones. Note that the maximum accuracy achieved for system S1 is only 4.5% below the reference system R2 using differential coefficients. This shows that three segmental coefficients capture almost 95% of the information contained in 13 differential coefficients. Thus, segmental coefficients can be more reliably estimated given sparse training data.

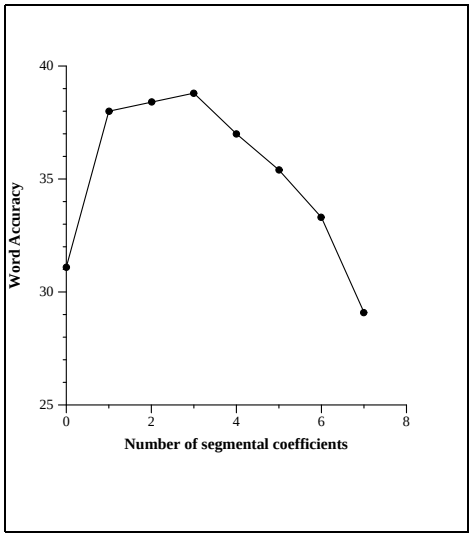

Figure 3: Word Accuracy for system S1

Finally, segmental coefficients were combined with mel-cepstral and differential coefficients yielding 27–dimensional feature vectors. This system (S2) outperforms R2 by 7.1% at the cost of one additional coefficient.

5. CONCLUSIONS

It was shown that incorporating segmental properties of speech cepstra significantly increases word recognition accuracy of speaker–independent HMM word recognisers. The sum of insertion and deletion errors drops from 24.5% to 23.2% which shows that less segmentation errors are made. The 1.6% reduction of substitution errors shows that segmental coefficients also contribute to the discrimination of otherwise confusable sounds. The results also indicate that delta cepstral coefficients do not capture all relevant transitional properties of phones.

Currently, experiments are conducted on a larger corpus to investigate the application of segmental coefficients in a multiple–mixture context–dependent HMM recogniser.

6. REFERENCES

1. Ralph Algazi et al. Transform representation of the spectra of acoustic speech segments with applications–I: General approach and application to speech recognition. *IEEE Trans. Speech and Audio Processing*, 1(2):180–194, April 1993.
2. Claude Barras et al. HMM based acoustic–phonetic decoding with constrained transitions and speaker topology. In Antonio J. Rubio Ayuso, editor, *NATO–ASI Workshop on New Trends in Speech Recognition*, Bubion, Spain, June 1995. Springer–Verlag.
3. Sadaoki Furui. Speaker–independent isolated word recognition using dynamic features of speech spectrum. *IEEE Transactions on Speech and Audio Processing*, 34(1):52–59, February 1986.
4. Xuedong Huang et al. The SPHINX–II speech recognition system: An overview. Technical Report CMU–CS–92–112, School of Computer Science, Carnegie Mellon University, January 1992.

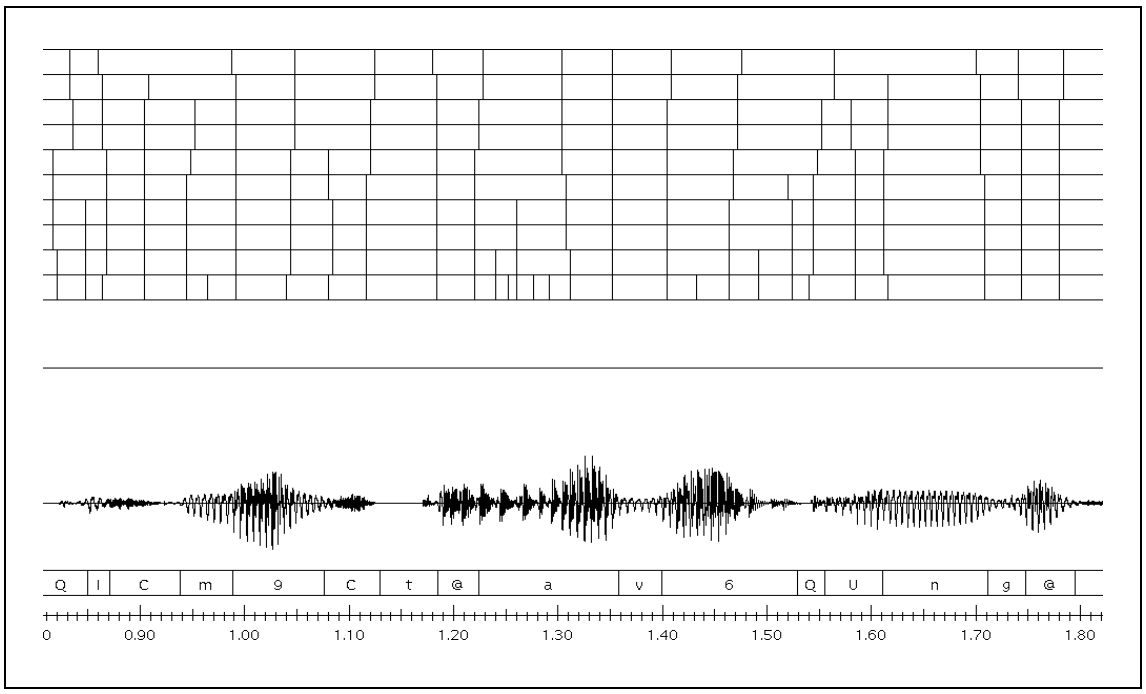

Figure 2: Example lattice with reference transcription

5. Kai Hübener and Andreas Hauenstein. Controlling search in segmentation lattices of speech signals. In *Proc. ECST'93*, pages 1763–1766, Berlin, Germany, September 1993.
6. Kai-Fu Lee. *Automatic Speech Recognition: The Development of the SPHINX System.* Kluwer Academic Publishers, Boston, Dordrecht, London, 1989.
7. Jeffrey N. Marcus. Phonetic recognition in a segment–based HMM. In *Proc. ICASSP'93*, pages 479–482, April 1993.
8. Carl D. Mitchell, Mary P. Harper, and Leah H. Jamieson. Using explicit segmentation to improve HMM phone recognition. In *Proc. ICASSP'95*, Detroit, May 1995.
9. Phil C. Woodland et al. Improving environmental robustness in large vocabulary speech recognition. In *Proc. ICASSP'96*, Atlanta (GA), May 1996.
10. Phil C. Woodland and Steve J. Young. The HTK tied–state continuous speech recogniser. In *Proc. ECST'93*, pages 2207–2210, Berlin, Germany, September 1993.