Speech Recognition at Multiple Sampling Rates

H.G. Hirsch[*]*, K. Hellwig, S. Dobler*

Ericsson Eurolab Deutschland
Nordostpark 12, 90411 Nuremberg, Germany
hans-guenter.hirsch@fh-niederrhein.de, {stefan.dobler, karl.hellwig}@eed.ericsson.se

Abstract

A feature extraction scheme is presented that analyzes speech signals sampled at different sampling rates. This will be needed in the future because of terminals in the telecom network that will transmit speech information also in the frequency region above 4 kHz.

A cepstral analysis scheme is applied in the frequency range up to 4 kHz to create a common set of acoustic parameters for all sampling rates. Additional parameters are determined describing the subband energy in the frequency region above 4 kHz.

As the major advantage of this feature extraction no individual recognizer has to be trained for each sampling frequency. It is shown with a recognition experiment that terminals and recognition systems can be combined without a remarkable loss in recognition performance with the terminal operating at a different sampling frequency than the recognizer has been trained on.

1. Introduction

The telecommunication network and the Internet are moving closer together and have already a lot of links between them. It might be that they will grow together completely in the future. This has already and will have even more influence on the application of recognition systems as they are applied in today's telecom network.

Up to now telephones operate in the spectral range up to 3.4 or 4 kHz. But the first wideband speech coding scheme has already been standardized in the field of mobile communication that analyses and transmits information in the spectral range up to 8 kHz. Such coding schemes offer an higher speech intelligibility and also a more natural speech perception by analyzing the broader spectral range. It is well known in the field of recognition that the analysis of a broader spectral range will lead also to higher recognition accuracy.

Thinking about the integration of PCs as terminals in the telecom network it is also feasible that speech will be sampled at other frequencies like 11 kHz. This sampling frequency is often used in the PC world derived from the CD sampling frequency of 44.1 kHz.

This shows that recognition systems will have the possibility in the future to process speech not only in the range up to 4 kHz. But such systems have to serve terminals that may operate at different sampling rates. This can be done most efficient with an extraction of acoustic features which does not create the need of training the recognizer separately for each sampling frequency.

A first feature extraction scheme has been presented and standardized by ETSI in the field of Distributed Speech Recognition (DSR) that supports sampling at 8, 11 and 16 kHz. The DSR approach is based on an implementation of the feature extraction in the terminal and the transmission of the acoustic parameters as data to a central recognizer in the network. But the design of this standardized analysis scheme creates the need of training the recognizer individually for each sampling rate to achieve highest performance.

In this paper we present an approach for the extraction of acoustic features that allows the operation of the feature extraction and the recognizer at any combination of different sampling frequencies without loosing recognition performance. The principal idea is presented first. Furthermore a practical implementation in a cepstral analysis scheme is shown with the presentation of recognition results for the TIDigits database.

2. Support of multiple sampling rates

A feature extraction scheme has been standardized by ETSI in the field of DSR that supports the sampling of speech at frequencies of 8, 11 and 16 kHz [1]. The analysis and recognition scheme is shown in figure 1 for the different spectral ranges.

[*] This author is now with the Niederrhein University of Applied Sciences in Krefeld, Germany.

Eurospeech 2001 (7th European Conference on Speech Communication and Technology)
Aalborg, Denmark, September 3-7, 2001

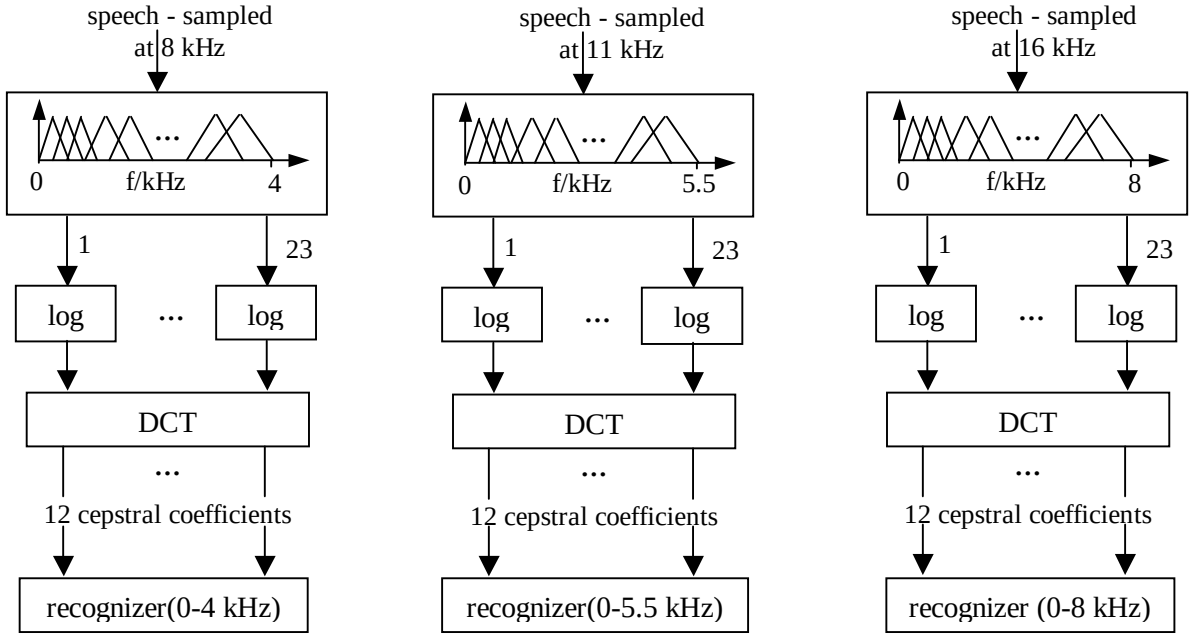

Figure 1: Feature extraction scheme (as standardized by ETSI) for multiple sampling rates

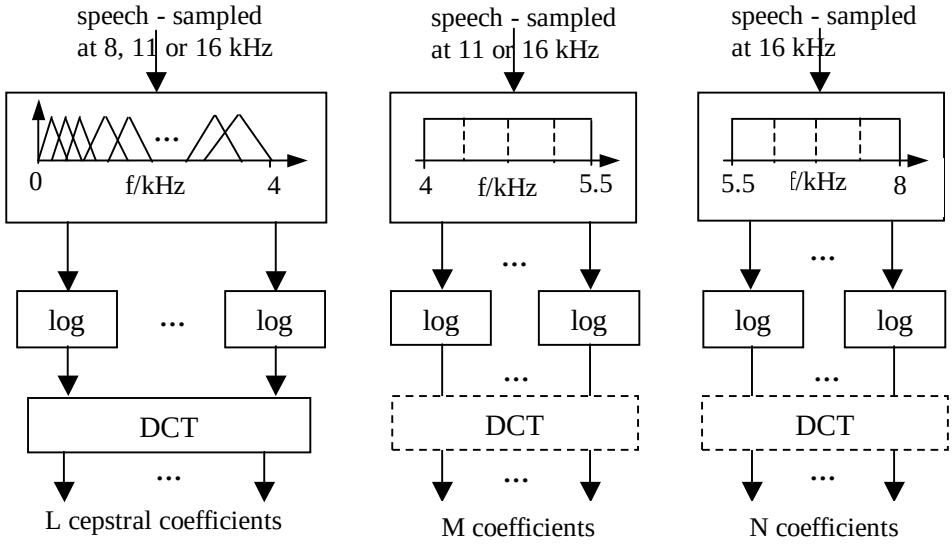

Figure 2: New feature extraction scheme for multiple sampling rates

12 cepstral parameters are extracted by applying a Mel filterbank to the different spectral ranges with a fixed number of 23 filters in the Mel frequency domain. Thus the triangular weighting functions for calculating the output of the Mel filterbank from the FFT spectrum look different and are placed at different center frequencies for each sampling rate. An individual recognizer has to be trained for each sampling frequency to achieve highest recognition performance.

Our approach is presented in figure 2 also supporting the extraction of acoustic features at sampling frequencies of 8, 11 and 16 kHz.

In general L cepstral coefficients are extracted by applying a Mel filterbank with a fixed number of filters in the spectral range up to 4 kHz at all sampling rates. The values of these coefficients are almost identical for the analysis of a speech segment at different rates.

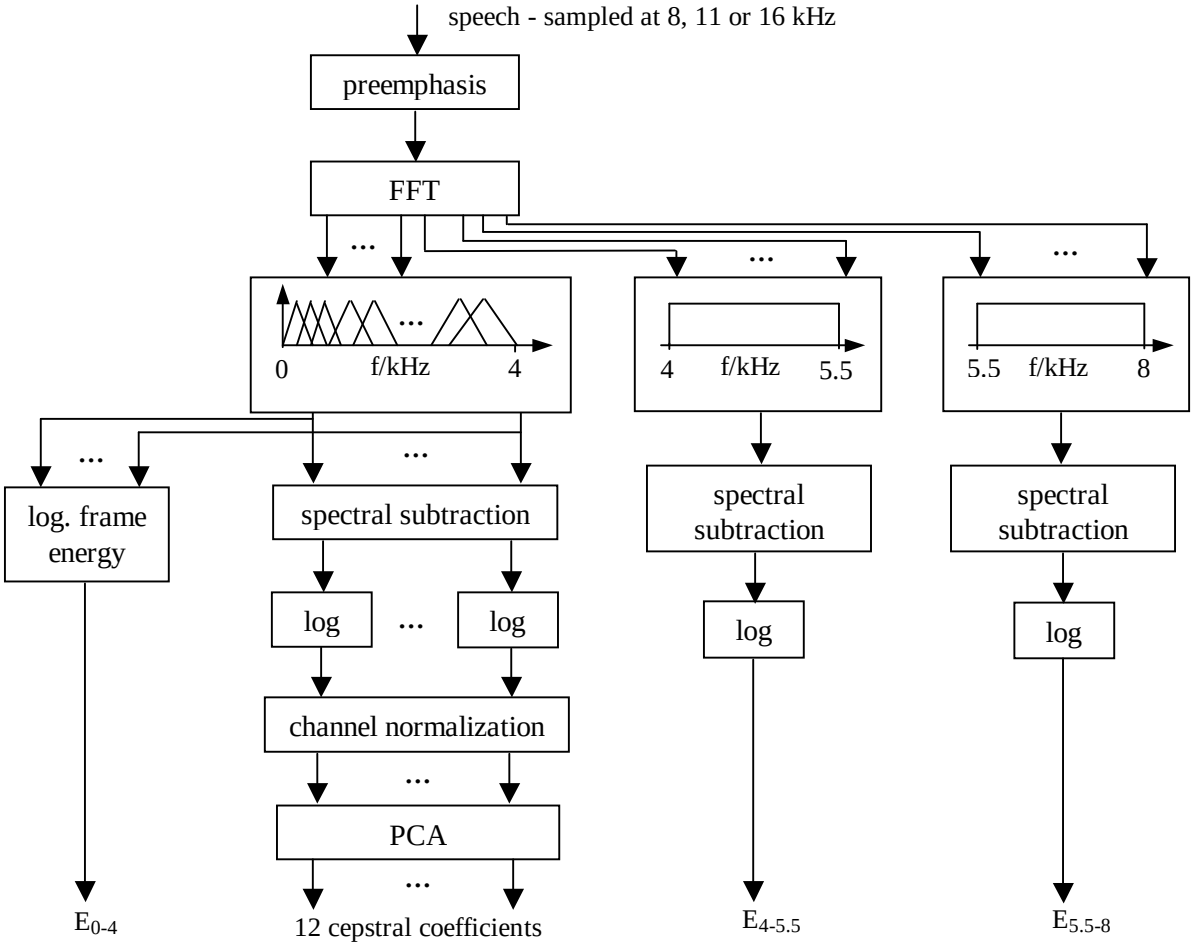

Figure 3: Analyzing speech at multiple sampling rates in a robust feature extraction scheme

As a consequence the preemphasis filter has to be adapted to the corresponding spectral range. At 8 kHz the usual preemphasis is applied by calculating the difference of consecutive samples with an preemphasis factor of 0.97. For the higher sampling rates FIR filters of length 12 have been designed that match the frequency characteristic of the usual preemphasis up to 4 kHz and creates a flat curve for frequencies above 4 kHz.

Further acoustic parameters are determined for the higher sampling rates in the spectral range above 4 kHz. This can be done by calculating the energies in several subbands in the range from 4 to 5.5 kHz and from 5.5 to 8 kHz as shown in figure 2. A simple solution could be just one additional parameter for each of the higher frequency regions. It is also feasible to calculate again cepstral coefficients for the higher spectral regions as indicated by the dashed DCT blocks in figure 2.

This approach creates a feature vector with a different number of components for each sampling rate but with a common set of cepstral coefficients for the spectral region up to 4 kHz. But this varying size of the feature vector can be easily handled by the recognizer. In case of operating a terminal at a higher sampling rate than the recognizer was trained on the additional components of the feature vector can just be neglected for the calculation of the emission probability. When operating a terminal at a lower rate than the recognizer was trained on the emission probability is just calculated for all components of the feature vector and neglecting the additional components in the states of the HMM models.

This approach creates the flexibility to combine terminals and recognizers that operate at different sampling rates.

3. Integration in a robust feature extraction

The basic idea is integrated in a robust feature extraction scheme as shown in figure 3.

Dependent on the sampling rate the preemphasis is individually done as described above. To analyze always speech segments of 25 ms duration a FFT of length 256 is applied at 8 kHz sampling and a FFT length of 512 is chosen for the higher rates. Three different sets of triangular weighting functions are needed for the 3 rates to create the Mel spectrum in the range up to 4 kHz. To achieve robustness against additive background noise, a spectral subtraction scheme is applied in the linear spectral domain. A channel normalization technique is applied in the logarithmic spectral domain to compensate the influence of unknown frequency characteristics.

One respectively two additional parameters are determined for the higher sampling rates describing the energies in the range from 4 to 5.5 kHz and from 5.5 to 8 kHz. A spectral subtraction is applied to those coefficients too. Finally the feature vector consists of 13 components at 8 kHz sampling rate and of 14 components at 11 kHz and of 15 components at 16 kHz.

4. Recognition results

The recognition performance is investigated on the TIDigits database [2] with a HTK (Hidden Markov Model Toolkit) based recognizer [3]. Three versions of all adult data are created by downsampling the data from the original 20 kHz to 16, 11 and 8 kHz. The intention is to create a set of HMMs for each sampling frequency. Then each of these recognizers can be used for the recognition of the designated TIDigits test data sampled at each of the three rates. Thus the complete matrix can be determined for all combinations of a terminal operating at one of the 3 sampling rates and a recognizer trained on data at one of the 3 rates.

The size of the feature vector is increased by a factor of 3 because of applying an LDA derived filtering [4] to each acoustic coefficient. This is comparable to calculating the Deltas and Delta-Deltas.

Whole word HMMs with 16 states per word and a mixture of 3 Gaussians per state are determined for each sampling frequency from the designated TIDigits training data (~8600 utterances). Thus 3 sets of HMMs are available. Details about the training can be found in [5]. The word error rates are listed in table 1 for all combinations of training the recognizer and analyzing the speech at individual sampling rates. All designated TIDigits test data (~8700 utterances) are fed into the recognizer.

HMMs trained on	Testing at 8 kHz	Testing at 11 kHz	Testing at 16 kHz
8 kHz	1.04	1.10	1.02
11 kHz	1.03	0.75	0.84
16 kHz	1.07	0.89	0.84

Table 1: Word error rates (%) at multiple sampling rates

First of all a gain can be seen when moving from 8 kHz to higher sampling frequencies and applying a recognizer trained at the same sampling rate. No further gain can be achieved when moving from 11 to 16 kHz. But this is not surprising because there does not exist much spectral information above 5.5 kHz for speech signals. Furthermore it can be seen that each terminal operating at one of the 3 sampling rates can be combined with each recognizer without a remarkable loss in recognition performance. This verifies the basic idea of our new concept. It will considerably reduce the effort of individually training a recognizer for each sampling frequency.

5. References

[1] "Speech processing, transmission and quality aspects (STQ); Distributed speech recognition; Front-end feature extraction algorithm; Compression algorithm", *ETSI standard document ES201108 V1.1.2*, 2000

[2] R.G. Leonard, "A database for speaker independent digit recognition, *ICASSP84*, Vol.3, p.42.11, 1984

[3] S. Young, D. Kershaw, J. Odell, D. Ollason, V. Valtchev, P. Woodland, "The HTK book – version2.2", Entropic, 1999

[4] Hermansky, H. and van Vuuren, S., "Data-driven design of Rasta-like filters", *Eurospeech97*, Rhodes, Greece, pp.409-412, 1997

[5] H.G. Hirsch, D. Pearce, "The Aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions", *ISCA workshop on automatic speech recognition*, Paris, France, 2000