SPEECH PARAMETERIZATION FOR AUTOMATIC SPEECH RECOGNITION IN NOISY CONDITIONS

Bojana Gajić

Department of Telecommunications,
Norwegian University of Science and Technology
7491 Trondheim, Norway
gajic@tele.ntnu.no

Kuldip K. Paliwal

School of Microelectronic Engineering,
Griffith University
Brisbane, QLD 4111, Australia
K.Paliwal@me.gu.edu.au

ABSTRACT

This paper is concerned with increasing the robustness of automatic speech recognition systems (ASR) against additive background noise, by finding speech parameters that are less influenced by changes in acoustic environments than the conventional ones.

Inspired by the good robustness of auditory based speech parameterization methods, we compare the steps involved with those in the conventional methods from the signal processing point of view. The use of dominant spectral frequencies is believed to be an important reason for the superior robustness of the auditory based methods.

A new speech parameterization method is described that is conceptually similar to auditory based methods, while retaining the low computational cost of the conventional methods. Evaluation on an ASR task has shown that the new method outperformed the conventional methods in presence of various background noises.

1. INTRODUCTION

State-of-the-art automatic speech recognition (ASR) systems are capable of achieving a very high recognition accuracy when tested in laboratory conditions. However, they usually experience a dramatic decrease in performance when used in real-world applications. One of the main reasons for such a behavior is presence of background noise in the testing environment that has not been observed during system training. This problem becomes especially important for ASR on mobile devices, as the acoustic environment is constantly changing and cannot be accounted for during system training.

One way to overcome this problem is to find a speech parameterization that is invariant to changing acoustic environments. The most commonly used speech parameters are based on the energy information derived from the short-term speech spectrum. However, the dominant spectral frequencies are less influenced by additive noise than the energy information. Thus, it is expected that the robustness of ASR systems could be improved if the dominant spectral frequencies are efficiently incorporated into speech parameter vectors.

The paper is organized as follows. It starts with an overview of ASR systems in Section 2, and describes the robustness problem with possible solutions in Section 3. Section 4 summarizes the main processing steps involved in conventional and auditory based speech parameterization methods and describe a new method that combines the advantages of both classes of methods. An experimental study performed to compare the performance of the different parameterization methods on an ASR task in various acoustic environments is described in Section 5. Finally, the major conclusions are summarized in section 6.

2. THE ASR SYSTEM

The aim of automatic speech recognition (ASR) is to transform a given spoken utterance into the corresponding transcription. A block diagram of an ASR system is shown in Figure 1. Before the system can be used, it has to learn the characteristic speech patterns from a large speech database with accompanying transcriptions. A set of stochastic models (hidden Markov models) is trained, each corresponding to one speech unit (for example phoneme). In addition, a lexicon is prepared to describe how the words are build up from the basic speech units, as well as a language model describing the relationship between words. The models, lexicon and language model are then used to determine the most likely transcription of an incoming spoken utterance.

The speech parameterization block is used to extract from the speech waveform the relevant information for discriminating between different speech sounds. The information is presented as a sequence of parameter vectors. This paper describes several different approaches to speech parameterization, and compares

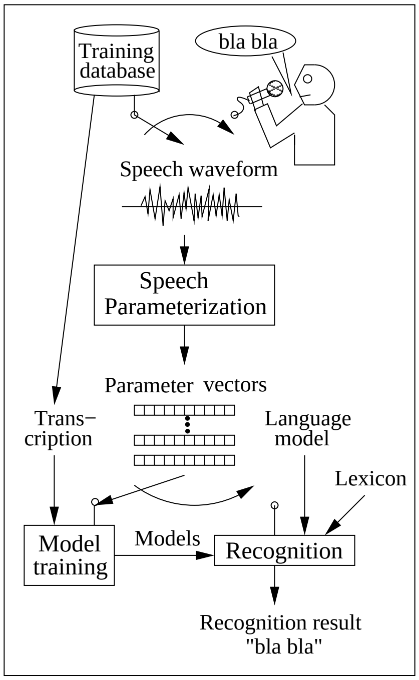

Figure 1: Block diagram of an ASR system

their performance on an ASR task in various noisy conditions.

3. THE ROBUSTNESS PROBLEM

Robustness of an ASR system is the system's ability to successfully deal with different aspects of variability in the speech signal. Some of the common variabilities that occur in speech signals are listed below:

- Pronunciation variations between speakers depending on speakers' voice characteristics, dialect, social class, etc.
- Pronunciation variations for a given speaker depending on mood, emotions, context, etc.
- Variations in the acoustic environment.
- Variations in the transmission channel.

A number of techniques have been proposed to increase the robustness of ASR systems. Nevertheless, it still remains a major obstacle for reliable use of ASR technology in many real-world applications. As the mobile hand-held terminals become more common, the robustness against variations in the acoustic environment becomes increasingly important. State-of-the-art ASR systems experience a dramatic performance degradation when the acoustic environment differs from the one observed in the training. In the following, we list the major classes of approaches for overcoming this problem.

Multiconditional training: The idea is to train a separate set of models for each background environment likely to occur during system use. For a given acoustic environment, the most likely set of models is then found and used during the recognition process.

Noise reduction: This approach is concerned with reducing the presence of noise in the speech signal before it is sent to the recognizer. When the models are trained in noise-free environments, this will reduce the mismatch between the input speech signal and the models. A most common approach is to apply noise spectral subtraction.

Model compensation and adaptation: Instead of modifying the speech signal to better comply with the models, in this approach the models are changed according to the statistical characteristics of the noise to better comply with the noisy speech.

Robust speech parameterization: The aim is to find such a speech representation that is invariant to changes of the acoustic environment. Note that this approach differs from the other approaches in that it does not require the knowledge of a particular acoustic environment during the use of the system. In the rest of this paper, we will focus on this approach.

4. SPEECH PARAMETERIZATION

This section starts with a summary of the major processing steps involved in conventional methods for speech parameterization. It proceeds by explaining the idea behind auditory based methods that have been shown to outperform the conventional methods in noisy conditions. The major differences between the two classes of methods are then explained from the signal processing point of view. At the end, a new parameterization method is described, that combines the advantages of both conventional and auditory based methods.

4.1. Conventional Methods

Conventional methods for speech parameterization are based on extracting the information from the short-term power spectrum of speech. The speech signal is divided into overlapping speech frames of 20-30ms length, as the speech signal can be regarded stationary on such a short intervals. The short-term power spectrum is estimated for each frame using either discrete Fourier transform (DFT), fast Fourier transform (FFT), filter bank analysis or linear prediction analysis. The resulting spectral representation is usually modified by applying some auditory motivated processing. At the end, it is usual to perform a decorrelation transformation, as this simplifies the recognition process.

Mel-frequency cepstrum coefficients (MFCC) are the most widely used speech parameters for ASR. Figure 2 illustrates the major processing steps involved in their computation. The short-term speech spectrum is estimated using FFT. It is passed through a filter bank consisting of overlapping triangular bandpass filters uniformly distributed along the perceptually based mel-frequency scale. The choice of the filter bank is motivated by the knowledge on human hearing. A vector of subband log-energies is then computed and sent to a discrete cosine transform (DCT) for decorrelation purposes. The resulting DCT coefficients, referred to as MFCC, serve as a final representation of the given speech frame.

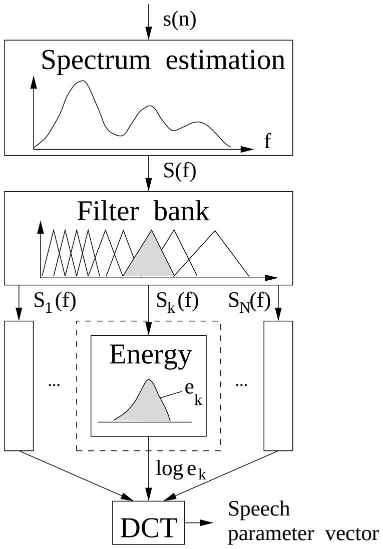

Figure 2: Illustration of MFCC computation

In the case of noisy speech, the subband energies get affected by noise, and the resulting speech representation differs from the one for clean speech. Thus, if an ASR system is trained on clean speech, and used in noisy conditions, the mismatch can cause a large performance degradation.

4.2. Auditory Based Methods

Humans have a fascinating ability to recognize speech in noisy acoustic environments. Thus, there is a belief that the robustness of ASR systems could be considerably improved by simulating the processes in human auditory system. However, not all the processes in human speech recognition are well understood, and auditory based methods for speech parameterization have to rely on some heuristics.

Probably the best known auditory based parameters for ASR are so called Ensemble Interval Histograms (EIH) [1]. In this paper, we will present a slight modification of these parameters referred to as Zero Crossings with Peak Amplitudes (ZCPA) [2]. These parameters have been shown to outperform both the EIH and all of the conventional parameterization methods in presence of additive noise. An illustration of the ZCPA method is shown in Figure 3. A frame of the given speech signal is passed through a filter bank of bandpass filters. The filtering is done in time domain. The resulting subband signals are sent to zero-crossing detectors. The interval between each pair of successive zero-crossings is measured together with the signal peak amplitude between the zero crossings. Then, the inverse intervals between successive zero crossings over all the subband signals are recorded in a histogram. Each histogram entry is weighted by the logarithm of the corresponding peak amplitude. Finally, the DCT is performed for decorrelation purposes.

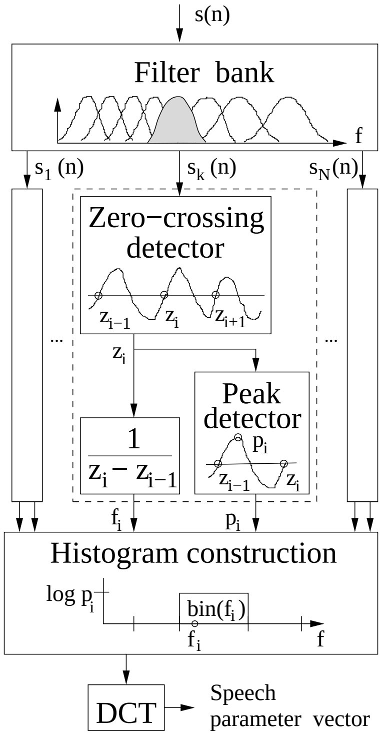

Figure 3: Illustration of ZCPA computation

Note that the ZCPA computation represents an alternative way of performing spectral analysis. The inverse intervals between successive zero-crossings represent the instantaneous dominant frequencies of the subband signal. The peak amplitudes, on the other hand, represent a measure of the instantaneous energy of the subband signal. The histogram bins containing the dominant frequencies are increased by the corresponding energy measures. Thus the resulting histogram represents an alternative representation of the signal spectrum.

While the MFCC is based only on the subband energy computation, ZCPA efficiently combines the energy and dominant frequency information. We believe that this difference can be a part of the explanation for the ZCPA's superior performance in noisy conditions. The dominant speech frequencies are much less affected by the presence of additive noise than the subband energy measures. Thus, incorporation of the dominant frequencies in the speech parameter vector can lead to increased robustness against additive noise. However, the ZCPA computation is prohibitively computationally expensive for use in practical ASR systems. This is due to time-domain processing and the need for heavy interpolation of the higher frequency subband signals in order to obtain a precise zero-crossing locations.

4.3. Subband Spectral Centroid Histograms

Motivated by the good noise robustness of the ZCPA parameters and the computational efficiency of the MFCC parameters, we searched for the possibility to design a new parameterization method, that would be more robust than MFCC, but have an acceptable computational cost. We believed that this task could be achieved by finding a more computationally efficient method for incorporating the dominant frequency information.

In [3] it has been shown that Subband Spectral Centroids (SSC) are closely related to the dominant speech frequencies. Using SSC as additional features to MFCC has been shown to increase the robustness of the ASR systems against additive noise [3, 4, 5, 6, 7].

We proposed a new framework for combining the SSC and subband energies through the construction of Subband Spectral Centroid Histograms (SSCH) [8, 9]. An illustration of the processing steps involved in the SSCH computation is shown in Figure 4. The speech power spectrum is estimated using FFT, and filtering is performed in the frequency domain to produce a number of subband signal. This part of the processing is analogue to the MFCC method. The dominant frequency of each subband signal is estimated by the subband centroid. In addition, a subband energy measure is computed similarly as for the MFCC method. The dominant frequency and energy information over all the subbands are combined in a single histogram in the same way as for the ZCPA method. Finally, the DCT is performed for decorrelation purposes.

This method uses the same conceptual information as the ZCPA method. However, note that the dominant frequencies are now estimated from the short-term power spectrum. This is a disadvantage in noisy conditions, as the spectrum itself is corrupted by noise. On the other hand, the fact that the processing is done in the spectral domain dramatically reduces the computational cost compared to ZCPA. It is now in the same order as for the MFCC computation.

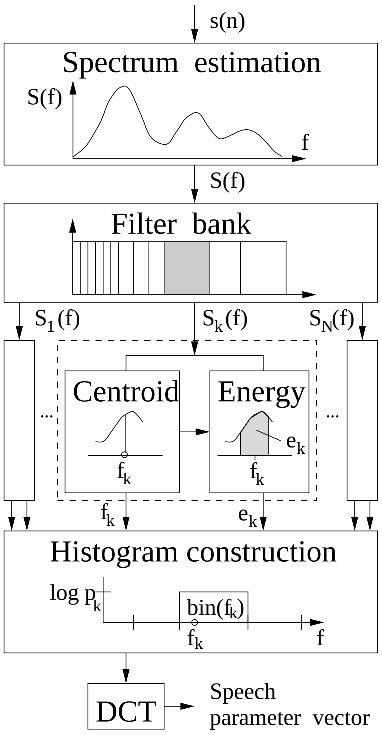

Figure 4: Illustration of SSCH computation

5. EXPERIMENTAL STUDY

This section describes an experimental study performed to compare the performance of the described methods on an ASR task in various background conditions.

5.1. Task and Database

The methods were evaluated on the ISOLET Spoken Letter Database [10] down-sampled to 8 kHz. The database consists of English letters spoken in isolation recorded in a quiet room. Two repetitions of each word were recorded for each speaker. Utterances from 90 speakers were used for training, while utterances from 30 speakers were used for evaluation. Although the vocabulary consisting of 26 English letters is rather small, this is not a simple recognition task, since the vocabulary words are very short and highly confusable.

Noisy speech was artificially created by adding to the original test set four different noise types at four different signal-to-noise ratios (SNR). Those are:

white Gaussian noise, factory noise, car noise and background speech. The last three noise types were taken from the NOISEX database, where they were referred to as factory1, volvo and babble noise respectively. A segment of the noise file equal to the length of the speech file was randomly extracted and added to the speech file at the required SNR. SNR was computed as the ratio between the maximal frame energy of the speech file, and the average energy of the noise segment. This way of computation makes SNR independent of the duration of the surrounding silence in the speech files.

Model training and recognition was performed using speech recognition toolkit HTK [11]. One hidden Markov model (HMM) with five states and five Gaussian mixtures per state was trained for each vocabulary word.

5.2. Choice of Free Parameters

In the following we summarize the most important parameters involved in MFCC, ZCPA and SSCH computation.

MFCC: Frame length was set to 25 ms. The filter bank consisted of 24 overlapping triangular filters uniformly spaced along the mel-frequency scale. 12 DCT coefficients were used. This is the standard parameter setting for the MFCC computation. It has not be optimized on the particular task.

ZCPA: The filter bank consisted of 20 bandpass FIR filters linearly spaced on the bark-frequency scale (perceptually based frequency scale similar to the mel-frequency scale), with bandwidths equal to 2 Bark. The filters had order 61, and were designed using the windowing method. Frequency dependent frame lengths equal to $20/f_c$ were used, where f_c is the center frequency of the corresponding bandpass filter. The number of histogram bins was 26. Number of DCT coefficients was 12.

SSCH: Frame length was set to 25 ms. The filter bank consisted of 65 rectangular filters. In the low frequency range, filter bandwidth was 300 Hz and the filters were linearly spaced along the frequency scale. In the high frequency region, filter bandwidth was 2 Bark and the filters were linearly spaced along the bark-frequency scale. 12 DCT coefficients were computed from 26 histogram bins.

Delta and delta-delta parameters were computed in addition to the static parameters for all of the methods, resulting in 36-dimensional parameter vectors.

5.3. Experimental Results

Table 1 shows the results of the evaluation of MFCC, SSCH and ZCPA parameterization methods on both clean and noisy versions of the ISOLET database. Model training was performed using clean speech. The recognition performance was measured in terms of word accuracy.

Table 1: Word accuracy for different parameterization methods in various acoustic environments

a) White Gaussian noise

Param.	SNR [dB]				
method	clean	25	20	15	10
MFCC	89.55	76.86	67.44	48.33	17.44
SSCH	87.24	78.91	70.58	57.69	38.21
ZCPA	85.19	76.68	71.28	62.37	48.08

b) Car noise

Param.	SNR [dB]				
method	clean	20	10	0	-5
MFCC	89.55	81.15	69.87	46.54	22.37
SSCH	87.24	86.92	86.15	81.35	72.69
ZCPA	85.19	85.19	82.31	73.27	61.47

c) Factory noise

Param.	SNR [dB]				
method	clean	20	15	10	5
MFCC	89.55	78.78	66.99	46.35	21.09
SSCH	87.24	79.36	71.79	54.36	35.96
ZCPA	85.19	78.65	71.67	59.87	37.31

d) Background speech

Param.	SNR [dB]				
method	clean	20	15	10	5
MFCC	89.55	73.14	57.56	39.04	22.18
SSCH	87.24	73.46	60.38	40.51	23.01
ZCPA	85.19	76.22	67.44	50.19	30.32

Looking at the results in Table 1, we see that MFCC performs best on clean speech. However, even in presence of only a small amount of noise, the situation changes completely, and MFCC becomes the worst of the three methods. This confirms the lack of the robustness of MFCC parameters.

SSCH is significantly more robust than MFCC for all the noise types. The improvement is largest for car noise, and smallest in presence of background speech. The relatively poor performance in presence of background speech is probably due to the existence of speech-like spectral peaks in the background signal.

SSCH even outperforms the ZCPA in the case of car noise, while ZCPA is more robust in presence of the other noise types. However, it is important to note that ZCPA cannot be used in place for SSCH in practical applications, due to its prohibitive computational cost.

6. CONCLUSIONS

In this paper, we addressed the robustness problem of the ASR systems against additive background noise. One way of overcoming this problem is to find a speech parameterization that is less influenced by additive noise than the conventional parameters.

We compared the steps involved in conventional and auditory based methods, and concluded that the superior performance of the auditory methods can be explained by the incorporation of the dominant spectral frequencies into parameter vectors.

A new speech parameterization method was described that computes the dominant spectral frequencies in a more efficient way, from the short-term spectrum of speech. Also this method outperformed the conventional methods in noisy conditions, confirming the importance of utilizing the dominant spectral frequencies for increasing the robustness of the ASR systems.

7. REFERENCES

[1] O. Ghitza, "Auditory models and human performance in tasks related to speech coding and speech recognition," *IEEE Trans. on Speech and Audio Processing*, vol. 2, pp. 115–132, January 1994.

[2] D.-S. Kim, S.-Y. Lee, and R. M. Kil, "Auditory processing of speech signals for robust speech recognition in real-world noisy environments," *IEEE Trans. on Speech and Audio Processing*, vol. 7, pp. 55–69, January 1999.

[3] K. K. Paliwal, "Spectral subband centroid features for speech recognition," in *Proc. ICASSP*, vol. 2, pp. 617–620, May 1998.

[4] S. Tsuge, T. Fukada, and H. Singer, "Speaker normalized spectral subband parameters for noise robust speech recognition," in *Proc. ICASSP*, May 1999.

[5] D. Albesano, R. D. Mori, R. Gemello, and F. Mana, "A study of the effect of adding new dimensions to trajectories in the acoustic space," in *Proc. EUROSPEECH*, vol. 4, pp. 1503–1506, September 1999.

[6] R. D. Mori, D. Albesano, R. Gemello, and F. Mana, "Ear-model derived features for automatic speech recognition," in *Proc. ICASSP*, 2000.

[7] E. Gjelsvik, "Modification of front-end processing for robust speech recognition." Diploma thesis, Norwegian University of Science and Technology, June 1999.

[8] B. Gajić and K. K. Paliwal, "Robust feature extraction using subband spectral centroid histograms," in *Proc. ICASSP*, May 2001.

[9] B. Gajić and K. K. Paliwal, "Robust parameters for speech recognition based on subband spectral centroid histograms," in *Proc. EUROSPEECH*, September 2001.

[10] R. A. Cole, Y. K. Muthusamy, and M. Fanty, "The ISOLET spoken letter database," Technical report CSE 90-004, Oregon Graduate Institute of Science and Technology, Beverton, OR, USA, March 1990.

[11] S. Young, D. Kershaw, J. Odell, D. Ollason, V. Valtchev, and P. Woodland, *The HTK Book*. Entropic, 1999.