BINAURAL BARK SUBBAND PREPROCESSING OF NONSTATIONARY SIGNALS FOR NOISE ROBUST SPEECH FEATURE EXTRACTION

Mike Peters

BMW AG
Research and Development Center
80788 Munich, Germany
e-mail: mike.peters@bmw.de

ABSTRACT

A two channel approach to noise robust feature extraction for speech recognition in the car is proposed. The coherence function within the Bark subbands of the Mel-Frequency-Cepstral-Transform is calculated to estimate the spectral similarity of two statistic processes. It is illustrated how the coherence of speech in binaural signals is used to increase the robustness against incoherent noise. The introduced preprocessing method of nonstationary signals in two microphones results in an *additive correction term* of the Mel-Frequency-Cepstral-Coefficients.

INTRODUCTION

The speech recognition in a noisy environment requires noise robust speech features. Since the internal Hidden-Markov-Models (HMM) of the speech recognizer are often trained by clean speech data, a discrepancy between the training data and the actual test data exists. Therefore, without preprocessing of the input signals or forced noise adaption of the trained speech models a degradation of the recognition accuracy is unavoidable. Basically[1], the Mel-Frequency-Cepstral-Coefficients as speech model based homomorphic deconvolution technics have been proved to be suitable for voice recognition in the past ten years.

Various speech enhancement and noise reduction systems in speech recognizer front-ends have been introduced. Often, noise reducing algorithms assume quasi uncorrelated stationary properties of speech and interfering noise (Wiener filtering) or use acoustic channel models (e.g. microphone arrays, adaptive noise compensation) and apply the knowledge of human speech perception to enhance the speech signal quality.

Microphone array and multiple channel solutions have been developed lately, but for the MFCC-Hidden Markov Model (HMM) based speech recognizer only one microphone approaches for the speech feature extraction are common.

This paper describes a signal preprocessing method based on a combination of the speech production models and *binaural* speech perception models. To estimate the relevant speech signal, the proposed algorithm takes advantage of the properties of speech and noise and their acoustic propagation in the car. It calculates the noise robust speech features *within* the MFCC's by application of a modified additive correction term received from the internal Bark subband coherence estimation of *two* input signals.

TWO CHANNEL APPROACH (MFCC-2)

The coherence function C_{xy} describes the correlation of two stationary signals *x(t)* and *y(t)* in the frequency domain:

$$C_{xy}(\Omega) = \frac{|S_{xy}(\Omega)|^2}{S_{xx}(\Omega)S_{yy}(\Omega)} \quad (1)$$

Following examples show the estimated coherence function of recorded speech and noise at different microphone positions.

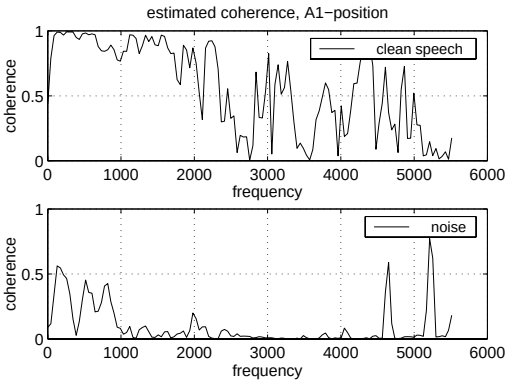

Figure 1: estimated coherence, A1-position

1. besides e.g. LPC features, extended or modified MFCC's

Figure 1 displays the estimated coherence of clean speech and vehicle noise with both microphones attached to the A-panel at drives side having a distance of appr. 30 cm (A1-position). In figure 2 each microphone has been positioned at the center of both sunvisors (S2-position).

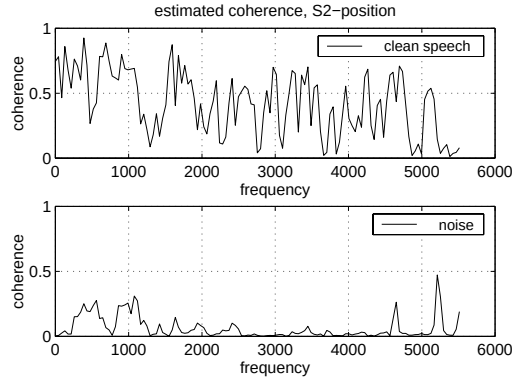

Figure 2: estimated coherence, S2-position

As shown in Figure 1 and 2, the coherence of noise at lower frequencies complicates the separation of speech and noise. The coherence separation method is not suitable at these frequencies and modification of the algorithm has to be applied. If the two microphones are close to the speaker then the speaker can be considered as concentrated source, while the interfering noise often has the properties of an expanded source. Taking into account the decreased coherence of clean speech at bigger microphone distances, an optimal microphone position (fig. 1) should be found. Hence it implies that within the radius r_{min} multiple path propagation of the signal can be neglected. To realize this behavior the minimum speaker microphone distance r_{min} referring to [3] is:

$$r_{min} = \sqrt{\frac{\gamma_s \gamma_e \bar{\alpha}_{Ab} A}{16\pi}} \tag{2}$$

wherein γ_s, γ_e are the aperture values of the speaker and the microphone and α_{Ab} is the absorption factor with A as absorption surface.

INTEGRATED MFCC CORRECTION

Assuming the one-channel clean speech signal *s(t)* is interfered by additive noise *n(t)*, further *n(t)* and *s(t)* are independent so that:

$$x^{(m)}(t) = [s^{(m)}(t) + n^{(m)}(t)]w(t) \tag{3}$$

where *(m)* is the index of the time frame and *w(t)* is e.g. a Hanning windowing function. The signal $x^{(m)}(t)$ is filtered by an auditory [4] Bark bank $\{b_{(i)}\}$ with i=1..K overlapping filters having triangular or gaussian shape and an equally spaced bandwidth of *1 bark = 100 mel*. The energy of the outcoming signal from each Bark filter $b_{(i)}$ is calculated. Hence the mel-vector $\{M^{(m)}\}$ of the $(m)^{th}$ frame becomes:

$$\{M_x^{(m)}\} = \sum_{i=1-\infty}^{K \quad \infty} \int \left| x(t) * b_{(i)}(t) \right|^2 dt \tag{4}$$

Applying the normalized discrete cosine transform (DCT) to *(2)* the mel-cepstral-coefficients $\{c_q^{(m)}\}$ are:

$$c_q^{(m)} = \sum_{i=1}^{K} log M^{(m)}(i) cos \frac{\pi q(2i+1)}{2K} \tag{5}$$

Since the speech and noise have nonstationary character the appropriate coherence function *(1)* is not defined. To estimate this function $C_{xy}^{(m)}$ within a time frame *(m)* following approach has been chosen:

$$C^{\circ(m)}_{xy} = \frac{\left| M_{xy}^{(m)} \right|^2}{M_x^{(m)} M_y^{(m)}} \tag{6}$$

where $C^{\circ}_{xy}{}^{(m)}$ is the coherence peak of the two input signals within a Bark subband. Applying Parseval's theorem to *(6)* and using *(4)* it follows the i^{th} Bark subband coherence becomes:

$$C^{\circ(m)}_{xy}(i) = \frac{\int \left| X^{(m)} B_i Y^{(m)*} B_i \right|^2 df}{\int X^{(m)} B_i X^{(m)*} B_i df \cdot \int Y^{(m)} B_i Y^{(m)*} B_i df} \tag{7}$$

To avoid residual error effects caused by peak estimation of C°_{xy} following recursive nonlinear 3-median filtering has been chosen:

$$C_{xy}^{(m)}(i) = median\{C_{xy}^{(m-2)}(i), C_{xy}^{(m-1)}(i), C^{\circ(m)}_{xy}(i)\} \tag{8}$$

The calculated coherence weight vector $C_{xy}^{(m)}(i)$ *reshapes* the Bark filter bank according to the subband coherence of the two audio signals and corrects the Melvector from (4) to:

$$\{M_{NR}^{(m)}\} = \sum_{1}^{K} M_x(i) C_{xy}^{(m)}(i) \tag{9}$$

Hence, inserting (5) and (9) into (5) the noise robust MFCC's $c_{NRq}^{(m)}$ are calculated by:

$$c_{NR_q}^{(m)} = c_q^{(m)} + \sum_{i=1}^{K} log[C_{xy}^{(m)}(i)] cos \frac{\pi q(2i+1)}{2K} \tag{10}$$

The second term in *(10)* may be considered as a additive cepstral correction offset to trained clean speech features within the frame *(m)*.

IMPLEMENTATION

Fig. 3 presents the implemented MFCC system. The time frames *(m)* have the length of 256 samples according to a sampling frequency of 11025 Hz. The frames overlap with 50% and are weighted by a Hanning window *w(t)*. As proposed in [3] the two framed input signals are highpass filtered to prevent low frequencies from disturbing the Bark subband estimation. Car specific low frequency noise and its propagation causes coherence estimation errors as indicated in Fig.1, 2. and effect a bad algorithm performance. An additional preemphasing filter results in a better delay compensation of the two microphone paths and a increased overall quality of the coherence estimation.

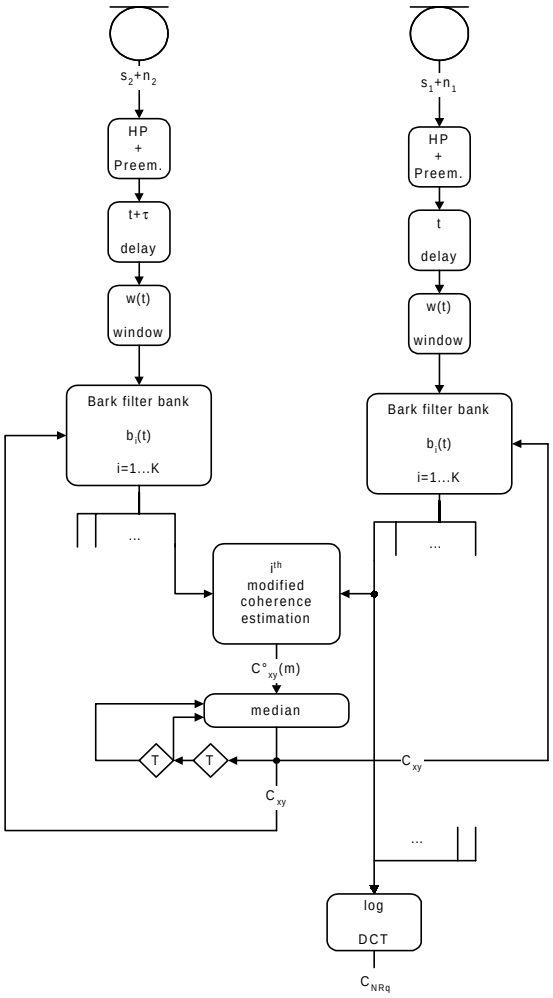

Figure 3: Implemented MFCC-2 system

An adaptive signal delay compensation [6] preserves the MFCC extraction from different propagation times of the speech into the two microphones. The applied Hanning windowing filter absorbs the effects of discrete time processing of speech signals. The Bark bank utilizes K=27 overlapping and triangular shaped Bark subband filters with an equally spaced bandwidth of 1 bark = 100 mel. The modified coherence estimation is median filtered. The output C_{xy} reshapes the appropriate Bark filter. The log-DCT extracts the noise robust MFCC's. Therefore, there is no need for retraining of the already trained speech HMM-model's. This is a big advantage.

EXPERIMENTS

To evaluate and test the proposed processing method following experimental environments have been used: The clean speech data (5 speaker) and noise signals have been acquired at different microphone positions in a vehicle. The noise signals represent various driving and road conditions. An additional white noise source has been chosen to evaluate the overall correction performance. Clean signal and noise pattern are mixed to build different SNR levels. As shown in figure 4, three different noise situations are used to test the recognition robustness for car environmental conditions.

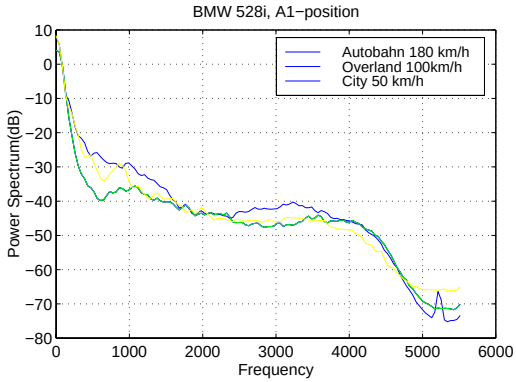

Figure 4: PSD of vehicle noise

A standard HMM speech recognizer for isolated words and a speech recognizer for continuous connected digits are used to evaluate the overall performance . The MFCC extraction has been modified as specified. The speech recognizer is trained with speech pattern acquired from approx. 1000 persons. The test set of isolated words consists of 200 German city names. The connected digits are spoken in digit blocks with a total length of a 500 phone numbers.

To compare the overall performance of this proposed processing method, an alternative one channel spectral subtraction algorithm and a two channel noise compensation algorithm are investigated. The spectral substraction incorporates an auditory speech enhancement system as introduced in [5]. The mixed noise and clean speech signals are processed and the output signal is sent to an unmodified HMM speech recognizer, using the same pattern recognition core.

The two channel noise compensation is taken from a system as proposed in [7]. Again, the processed signal mixture is sent to an unmodified HMM speech recognizer.

TEST RESULTS AND CONCLUSIONS

Figure 5 illustrates the averaged error probability received by comparison of three different noise sources with clean speech. It follows that the position of the microphone[2] has more influence on the recognition reliability then the noise source.

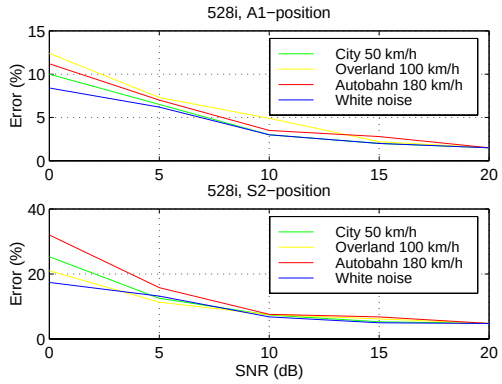

Figure 5: recognition reliability for different mic-positions

As shown in figure 5, the recognition rate for stationary noise is higher than for noise generated by rapidly changing driving conditions. However, the recognition reliability decreases for driving noise at 180 km/h speed while wind and tire noises are significant. It indicates that the robustness of speech recognition using this corrected feature extraction depends on the degree of correlation between noises in each channel. Figure 6 illustrates the overall performance for different preprocessing methods compared to an unprocessed speech input to a HMM recognizer.

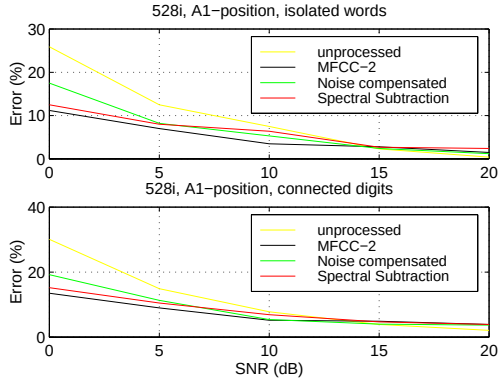

Figure 6: comparison overall performance

2. here: A1-two microphones at A-panel, S2- one microphone on both sunvisors

Hence it follows, the overall performance of word isolated recognition is better than one for connected digits. Especially at low SNR the performance of the proposed algorithm increases compared to the others. Another interesting point is that the error probability of the two channel noise compensation not as good as those for one channel spectral substration at higher SNR. Compared to the noise compensation and the spectral substraction the proposed processing method the improvement in noise robustness was approx. 10% and almost 15%, respectively.

SUMMARY

In this paper a preprocessing method for noise robust MFCC feature extraction are introduced. This algorithm utilizes the coherence of speech in a realistic vehicle environment for an estimation of the additive MFCC-correction term. The experimental results applying a modified HMM recognizer are summarized as follows: The MFCC-2 processing improves the performance and noise robustness both realistic and white noise. Compared to established preprocessing method, the performance increases about 10 to 15% which allows to neglect additional costs for a second microphone and codec and approx. 10% higher processing complexity.

REFERENCES

[1] M. Boden, K. Rateitschek: “Noise robust speech recognition based on a binaural auditory model”, Proc. of Workshop on the Auditory Basis of Speech Perception, pp. 291-252 (1996)

[2] S. Kajita, K. Takeda, F. Itakura: “A binaural speech processing Method using subband-cross-correlation analysis for noise robust recognition”, ICASSP 1997, pp.1243 -1247

[3] R. Martin: “Freisprecheinrichtungen mit mehrkanaliger Echokompensation und Störgeräuschreduktion”, 1.Aufl. Aachen, Verlag der Augustinus Buchhandlung, 1995, pp. 11-88

[4] E. Zwicker: “Psychoakustik”, Berlin, 1992

[5] H. Reininger, C. Kuhn: “Signalverbesserung durch gehörgerechte Spektrale Substraktion”, Proceedings of 9. Aachener Kolloqium “Signaltheorie”, Bild und Sprachsignale, Aachen 1997

[6] Y. Bar-Shalom, F. Palmieri, A. Kumar, H. Shertukde, “Analysis of Wide-Band Cross Correlation for Time-Delay Estimation”, IEEE Trans. on Signal Processing, Vol 41, No.1 pp.385-387, January 1993

[7] S. V. Vaseghi: “Advanced Signal Processing and Digital Noise Reduction”, Wiley Teubner N.Y. 1996