ROBUST SPEECH RECOGNITION WITH MULTI-CHANNEL CODEBOOK DEPENDENT CEPSTRAL NORMALIZATION (MCDCN)

Sabine Deligne and Ramesh Gopinath

IBM T.J. Watson Research Center, P. O. Box 218,
Yorktown Heights, NY 10598
deligne@us.ibm.com

ABSTRACT

In this paper, we address the issue of speech recognition in the presence of interfering signals, in cases where the signals corrupting the speech are recorded in separate channels. We propose to combine a trivial form of filtering with MCDCN, a Multi-channel version of the Codebook Dependent Cepstral Normalization, where the cepstra of the noise are estimated from the reference signals. We report on recognition experiments in a car where the speech signal is corrupted by radio talks or CD music played the car speakers. Our approach allows relative word error rate reductions in the range of 70-90% compared to a no-compensation baseline, at a relatively low computational cost.

1. INTRODUCTION

Robustness in the presence of noise, and more generally in the presence of interfering signals, is a crucial issue for speech recognition to work in a real-world environment. In cases where the signal interfering with the speech is stationary and where its characteristics are known in advance, robustness issues can, to a certain extent, be addressed during the training of the speech recognition system. However, in most applications, the signal corrupting the speech is neither known in advance nor stationary (for example, music or speech from competing speakers). Such cases cannot be handled by devising special training schemes and they require the use of on-line adaptation algorithms. In this paper, we address the case where recordings of the interfering signals are available in separate channels. These signals are called the reference signals. This occurs for example when the speech signal is corrupted by the sound emitted by a radio or a CD player (the reference signals are recorded at the outputs of the radio or CD player), in telephony applications where the speech prompt synthesized by the speech server interferes with the speech of the user (the reference signal is the recording of the prompt), or, when the speech signal is mixed with the speech of competing speakers (the reference signals are recorded from the microphones of the competing speakers). The general problem of removing unwanted signals from a desired signal by using reference signals is typically addressed with adaptive decorrelation filtering techniques [1]. In decorrelation filtering, the corrupted signal and the reference signal are assumed to be observed at the output of a linear system modeling the cross-coupling between the desired signal and the interfering signal. This linear system is assumed to be such that there is no leakage of the desired signal into the reference signal. Further assuming that the desired and interfering signal are uncorrelated, the linear system can be estimated unambiguously so that the desired signal can be recovered via inverse filtering. However adaptive decorrelation filtering suffers from some limitations in the context of speech recognition, especially in the context of embedded applications running with limited computational resources: (i) it performs in the waveform domain, on a sample basis, thus leading to a high computation rate, (ii) it involves an iterative scheme, hence some delay may occur before it converges towards an accurate estimate of the coupling system, especially in a non-stationary environment, (iii) its performance depends on the modeling accuracy of the coupling system (the length of the decorrelating filters needs to be hypothesized). In this paper, we present an approach especially designed to deal with a real time application constrained to run with low computational ressources. An inexpensive - and inaccurate - form of adaptive filtering, assuming a single-tap delay filter, is used to roughly align and scale the reference signal with the noisy speech. The aligned and scaled reference signal is then removed from the noisy speech in the cepstral domain by using our new algorithm derived from CDCN [3] and called MCDCN: Multi-channel Codebook Dependent Cepstral Normalization. As will be shown in this paper, MCDCN is advantageous as: (i) it allows to compensate for the loose modeling of the coupling system between the speech and the interfering signal by taking advantage of our knowledge of the clean speech distribution in the cepstral domain, (ii) it does so through the use of a codebook, the size of which can be adjusted to meet the desired balance between performance and computational complexity, (iii) it

performs on a frame basis, i.e. at a low computation rate compared to waveform techniques (every 165 samples with our 15ms system on 11kHz data, instead of every sample), (iv) it does not involve any iterative estimation scheme, thus further enabling a real time use.
In section 2, we present our multi-channel version of CDCN and in section 3 we give an overview of our noise removal scheme, including the preliminary alignment and scaling steps used in our experiments. In section 4, we report on speech recognition experiments in a car with either radio talks or CD music played by the car speakers at different sound levels.

2. MCDCN

MCDCN refers to a multi-channel version of CDCN that allows to compensate for non-stationary noise in cases where the source(s) of noise are recorded separately. In the standard CDCN framework, the desired speech signal is assumed to be first passed through a linear filter, which models the effect of the channel, and then corrupted with noise. In this paper, only the cepstral distortion caused by the noise is considered: the effect of the channel is assumed to be compensated for by the preliminary alignment and scaling explained in section 3. Thus, assuming additive uncorrelated noise, the relation between the power spectral densities of the clean speech, $P_y(f)$, of the noisy speech, $P_x(f)$, and of the noise corrupting the speech, $P_n(f)$, is:

$$P_y(f) \ = \ P_x(f) \ - \ P_n(f) \tag{1}$$

Note that equation 1 suggests that, given the corrupted speech and the noise observed in the reference channel, the spectrum of the clean speech could be recovered with spectral substraction. However our preliminary experiments tend to indicate that this would require to identify the cross-coupling system between the noise and the clean speech, so that the noise actually corrupting the speech can be estimated by filtering the noise in the reference channel. In [2], adaptive lattice-ladder filters are used to very accurately align the reference signal with the noise present in the corrupted speech. The aligned reference signal is then removed from the noisy speech by using a spectral substraction technique. The MCDCN technique presented in this paper allows to avoid the cost of an accurate alignment: the imprecision of our estimate of the corrupting noise, as well as the imprecision of the additive noise model, is compensated for by taking advantage of our *a priori* knowledge of the clean speech distribution in the cepstral domain. Besides, MCDCN does not require any empirical tuning whereas spectral substraction requires to define an adequate flooring of the cleaned spectrum. The relation between the cepstral vectors of the clean speech $y(t)$, the noisy speech $x(t)$ and the noise $n(t)$ can be expressed as [3]:

$$y(t) \ = \ x(t) \ - \ r(y(t), n(t)) \tag{2}$$

with r a non linear function of both the clean speech and the noise. Assuming MFCC vectors computed with a bank of Mel-filters followed by a Discrete Cosine Transform:

$$r(y(t), n(t)) \ = \ DCT \ \log(1 \ + \ e^{DCT^{-1} \ (n(t) \ - \ y(t))}) \tag{3}$$

where DCT and DCT^{-1} refer respectively to the Discrete Cosine Transform and to its inverse. Whereas in standard CDCN, the noise is estimated via an EM algorithm, we propose in MCDCN to compute the cepstra $n(t)$ of the noise from the reference signal which is assumed to be recorded in a separate channel. For lack of knowing the cepstra $y(t)$ of the clean speech, the function r, like in standard CDCN, is approximated with its expected value over y, given $n(t)$ and $x(t)$:

$$\hat{r}(y(t), n(t)) \ = \ E_{y(t)}[r(y(t), n(t)) \mid x(t), n(t)]$$

To simplify the computation, the function $r(y(t), n(t))$ is assumed to be a piece-wise constant function of $y(t)$. Therefore, assuming a codebook $C_{n_C} = \{ \ c_i \ \}_{i=1}^{n_C}$ of n_C cepstral vectors describing the acoustic space of the clean speech, the noise correction term is computed as:

$$\hat{r}(y(t), n(t)) \ = \ \sum_{i=1}^{n_C} \ p(c_i \mid x(t), n(t)) \ r(c_i, n(t)) \tag{4}$$

Assuming the Gaussian distribution $\mathcal{N}(y(t) \ ; \ \mu_i \ , \ \sigma_i^2)$ to model the distribution of the clean speech $y(t)$ given the codeword c_i, we approximate the distribution of the noisy speech $x(t)$, given c_i, with the Gaussian distribution $\mathcal{N}(x(t) \ ; \ \mu_i + r(c_i, n(t)) \ , \ \sigma_i^2)$. The posterior probability of the codeword c_i, given $x(t)$ and $n(t)$, is thus computed as:

$$p(c_i \mid x(t), n(t)) = \frac{\pi_i \mathcal{N}(x(t); \mu_i + r(c_i, n(t)), \sigma_i^2)}{\sum_{j=1}^{n_C} \pi_j \mathcal{N}(x(t); \mu_j + r(c_j, n(t)), \sigma_j^2)}$$

where π_i denotes the *a priori* probability of the codeword c_i. In [4], CDCN is used with a more refined estimation of the distribution of the noisy speech, inspired by the model combination framework. An estimate of the clean speech $\hat{y}(t)$ is computed as:

$$\hat{y}(t) \ = \ x(t) \ - \ \hat{r}(y(t), n(t)) \tag{5}$$

The computational cost of MCDCN is a linear function of the size n_C of the codebook. The codebook can thus be designed so as to find the desired balance between performance and computational complexity

3. OVERVIEW OF THE SYSTEM

The reference waveforms are first aligned (assuming a single-tap delay filter) and scaled against the noisy speech waveform. In our experiments, a speaker is talking in a car while the radio or the CD player are on. A microphone located on the visor of the car captures the speech corrupted by the signal emitted by the car speakers. The radio is a mono source of noise: its output is captured in one reference channel. The CD player is a stereo source of noise: the left and right outputs of the CD player are captured in two distinct channels. In the case of the mono source, the relative delay between the waveform in the reference channel and the waveform of the corrupted speech is estimated by detecting the maximum of the cross-correlation function between the two waveforms. The scaling factor between the amplitudes of the two waveforms is estimated by computing a mean value over segments of non speech samples from each waveform, and by taking the ratio of the two means. The estimated scaling factor is used to set the two input waveforms to approximately the same amplitude level. This preliminary step represents a simple form of filtering ; a more refined scheme like adaptive decorrelation filtering for example could be used instead, but at a much higher cost. In the stereo case, the signals from the left and right outputs of the CD player are aligned in turn against the waveform of the noisy speech, using the same cross-correlation technique as in the mono case. The left and right waveforms are then summed up so that even in the stereo case, only one reference signal is actually used when applying MCDCN. This reference signal is scaled against the noisy speech following the same technique as in the mono case. After the input waveforms have been aligned and scaled, cepstra are computed for each channel. The cepstra in the reference channel and in the noisy speech channel are used as estimates of respectively $n(t)$ and $x(t)$ in equations 4 and 5. Speech recognition is performed on the estimate $\hat{y}(t)$ of the cepstra of the clean speech obtained from equation 5.

4. EXPERIMENTS AND RESULTS

To collect the evaluation data, 20 subjects (10 males and 10 females) were given 50 sentences consisting of digit strings or command phrases. Each subject was asked to repeat the 50 sentences in a stationary car with the speakers playing either radio news or CD music (opera, DJ or jazz music) at 3 signal power levels: 60 dB, 70 dB and 80 dB in average, as measured by an SPL meter between the front seats at about lap level. The speech corrupted by the sound emitted by the car speakers was recorded with an AKG Q400 microphone located on the visor. Simultaneously, the signals at either the radio output or at the left and right outputs of the CD player were captured in separate channels. All the data were recorded at 22kHz and downsampled to 11kHz.

In the experiments presented here, the reference and speech channels are aligned by detecting the maximum of their cross-correlation function for shifts of up to 90ms. They are scaled by estimating the mean of the signal in each channel during the first 450ms of the recording (we assumed that there is no speech during the first 450ms of each sentence).

Speech recognition is performed with a reduced-size system especially designed for portable devices or automotive applications [5]. It consists of speaker-independent acoustic models (156 subphones covering the phonetics of English) with about 9,000 context-dependent gaussians (triphone contexts tied by using a decision tree), trained on a few hundred hours of speech (about half of these training data has either digitally added car noise, or was recorded in a moving car at 30 and 60 mph). The front end of the system uses 39 dimension cepstra [1] (12 MFCC + the energy + delta and delta-delta coefficients) from 15ms frames. MCDCN was applied by using codebooks of either 2, 4, 8, 16, 32, 64, 128 or 256 codewords. Each codebook was estimated by quantizing about 3,000 sentences of clean speech (recorded with the same microphone as the evaluation data) by assuming diagonal covariance matrices tied across all codewords. All codewords were assigned equal priors.

Table 1 shows the average word error rates (WER) obtained by decoding the noisy speech without using any compensation and by compensating with MCDCN with codebooks of size ranging from 2 to 256 codewords. The average is taken over all the speakers and all the interfering signals (radio and all music styles) at each of the three sound levels. With as few as 2 codewords, MCDCN allows a relative WER reduction of about 75% with the 60 and 70dB interferences, and 65% with the 80dB interferences. The best performance at 60 dB corresponds to an 82% WER reduction and it necessitates codebooks of at least 8 codewords. The best performance at 70 dB and 80dB corresponds to WER reductions of respectively 87% and 76% with codebooks of at least 32 codewords. Our experiments tend to indicate that the minimal number of codewords required to reach an optimal performance is in relation with the power level of the interfering noise: the louder the noise, the bigger the codebook needs to be. Our interpretation is that the approximation used in the CDCN framework, according to which $r(y(t), n(t))$ is a piece-wise constant function of $y(t)$, holds better at low levels of noise[2]. The histograms on figures 1 show the WER when the interfering signal is a) the opera, b) the DJ music, c) the jazz music and d) the radio news talk. The most confusing interference for the speech recognition system is the competing speech from the radio speaker, and then the DJ kind of music (note that the DJ

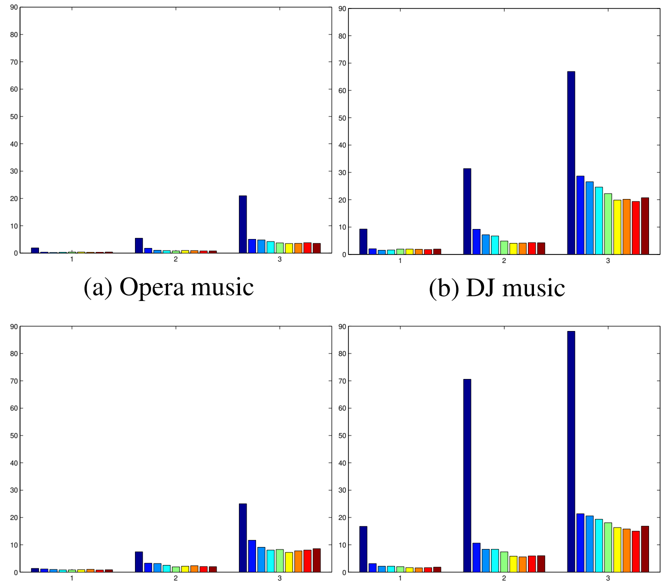
[1] cepstra of dimension 24 are used to apply the DCT in equation 3.

[2] Actually, it can be verified that, for a given $y(t)$, if $| n_1(t) | \leq | n_2(t) |$ then $| \frac{\partial r(y(t), n_1(t))}{\partial y(t)} | \leq | \frac{\partial r(y(t), n_2(t))}{\partial y(t)} |$.

tracks consisted of mainly rap music, i.e. somehow again a competing speech). The effect of the radio however is better compensated than the effect of the DJ music, possibly because the radio is a mono source and our simple alignment+scaling scheme can better approximate the channel effect than with a stereo source. In our experiments, the delays between the noisy speech waveform and the reference waveforms were found to be in the range 5 to 15 ms. We tried to alleviate the possible impact of mis-alignments by applying, within each channel, the cepstral averaging technique presented in [6]: each 15ms cepstrum is obtained by averaging 3 cepstra computed with 5ms shifts. It resulted in about 10% relative improvement at the lowest power level (0.9% versus 1.0%), but it hurt the accuracy at the loudest levels. This is consistent with the fact that mis-alignments are more likely to occur when the amount of interfering signal in the corrupted speech is small.

	60dB	70dB	80dB
no compensation	6.1	23.8	44.5
$n_C = 2$	1.5	5.5	15.6
$n_C = 4$	1.2	4.5	13.9
$n_C = 8$	**1.1**	4.1	12.8
$n_C = 16$	**1.2**	3.3	12.0
$n_C = 32$	**1.2**	**3.0**	**10.8**
$n_C = 64$	**1.1**	**3.0**	**10.8**
$n_C = 128$	**1.0**	**3.0**	**10.7**
$n_C = 256$	**1.2**	**3.0**	11.5

Table 1. Average WER over all speakers and interfering signals, for interfering signals at power level 60, 70 and 80dB, and for codebooks of various sizes



(a) Opera music (b) DJ music

(c) Jazz music (d) Radio news talk

Fig. 1. WER averaged over all speakers for each interfering source (opera,DJ,jazz,radio). The 3 groups of bars on each figure correspond to interfering signals at the power levels 1. (60dB), 2. (70dB) and 3. (80dB). The first bar at each sound level shows the WER when decoding without compensation, and the other bars show the WER when decoding with codebooks of increasing size: 2, 4, 8, 16, 32, 64, 128 and 256 from left to right.

5. CONCLUSION

A Multi-channel version of CDCN was proposed to compensate for the effect of interfering signals during speech recognition by using reference signals. Whereas adaptive filtering techniques focus on estimating the coupling system between the unwanted and desired signals, MCDCN approximates the effect of the unwanted signals in the cepstral space by taking advantage of our *a priori* knowledge of the clean speech distribution. As a result, it is computationally attractive compared to approaches that require adaptive filtering. In our experiments of ambient music removal in a car, word error rate reductions in the range of 70-90% were obtained.

6. REFERENCES

[1] E. Weinstein, M. Feder and A.V. Oppenheim, “Multichannel signal separation by decorrelation”, IEEE Transactions on Speech and Audio Processing, vol. 1, num.4, October 1993.

[2] P. Gomez-Vilda, A. Alvarez, R. Martinez, V. Nieto and V. Rodellar, “A hybrid signal enhancement method for robust speech recogntion”, Proceedings of the Workshop on Robust Methods for Speech Recognition in Adverse Conditions, Tampere, Finland, 1999.

[3] A. Acero and R.M. stern, “Environmental robustness in automatic speech recognition” Proceedings of ICASSP 90, 1990.

[4] M. Westphal and A. Waibel, “Model-combination-based acoustic mapping”, Proceedings of ICASSP 01, 2001.

[5] S. Deligne, E. Eide, R. Gopinath, D. Kanevsky, B. Maison, P. Olsen, H. Printz, J. Sedivy: “Low-resource speech recognition of 500-word vocabularies”, Proceedings of EUROSPEECH 01, 2001.

[6] S. Dharanipragada and R. Gopinath, “Smoothed cepstral trajectories for speech recognition” Proceedings of ICSLP98, 1998.