Robust speech recognition with Multi-Channel Codebook Dependent Cepstral Normalization (MCDCN)

Sabine Deligne, Satya Dharanipragada, Osamu Ichikawa

IBM T.J. Watson Research Center,
Yorktown Heights, NY 10598
* IBM Tokyo Research Laboratory, Japan
deligne@us.ibm.com

Abstract
Multi-channel Codebook Dependent Cepstral Normalization (MCDCN) aims at compensating for noise in the case where the source of noise can be recorded in a secondary channel. The general problem of removing unwanted signals (music, news talk...etc) from a desired signal (speech) when reference signals are available in separate channels is typically addressed with adaptive filtering techniques such as noise canceling. Unlike these conventional techniques, MCDCN operates directly on the cepstra of the signals, making it especially appropriate in the context of speech recognition. We compare the recognition accuracy obtained by enhancing speech with MCDCN or conventional noise canceling or a combination of both techniques. The comparison is made on real noisy speech data where the speech is corrupted by the signal played by a radio and a CD player in a car. In our speech recognition experiments, the WER with MCDCN are 3 to 5 times less than with noise canceling. Besides we show that the performance of MCDCN can be further enhanced by combining it with noise canceling.

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 most applications, the signal corrupting the speech is neither known in advance nor stationary (for example, music or speech from competing speakers) so that adaptive noise compensation algorithms are required to maintain the accuracy of the recognition. In this paper, we address the case where recordings of the interfering signals, the so-called reference signals, are available in separate channels. 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 barge-in enabled systems (e.g. the reference signal is the recording of the prompt synthesized at a speech server), 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 problem of signal separation with a reference signal is typically addressed with adaptive filtering techniques, such as noise canceling [1], which seek to estimate a filter to separate the primary and reference signals. Assuming that there is no leakage of the desired signal into the reference channel and that the desired and corrupting signal are uncorrelated, then the desired signal can be recovered unambiguously. However adaptive filtering suffers from some limitations in the context of embedded speech recognition applications with limited computational resources: (i) it performs in the waveform domain, on a sample basis, thus leading to a high computation rate, (ii) it usually involves a gradient descent algorithm, hence the need to tune learning parameters, the risk of instability and possibly the difficulty to track an accurate estimate of the coupling system in case of a time-varying environment, (iii) its performance depends on the modeling accuracy of the coupling system (e.g. length of the un-coupling filter).

In this paper, we present an approach especially designed to deal with a real time application constrained to run with low computational resources. 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 in a system with 15ms frame shift and 11kHz sampling rate, instead
of every sample), (iv) it does not involve any learning algorithm, thus further enabling a real time use.

The rest of the paper is organized as follows. In section 2 we review the noise canceling scheme used in our experiments. In section 3, we present our multi-channel version of CDCN. In section 4, we report on speech recognition experiments in a car with real noisy speech data where the corrupting signal is either talk-radio or CD music played by the car speakers at different sound levels. We compare the accuracy obtained with echo canceling, MCDCN and a combination of both techniques.

2. Adaptive Noise Canceling

Assume that we observe a primary sensor signal \( s_1 \) (e.g. mixture of speech and music captured by a microphone) and a secondary (or reference) sensor signal \( s_2 \) (e.g. music captured at the output of a CD/radio player). The most widely used approach to the two-channel signal separation problem in the case where there is no coupling of the desired signal (e.g. clean speech) into the reference sensor was suggested by Widrow et al. [1]. It is assumed that the signal interfering with the desired signal is coupled into the primary sensor through an unknown filter whose input is the signal \( s_2 \) measured by the secondary sensor. The unknown filter is identified by minimizing the average power of the reconstructed signal. Minimizing the average power corresponds to estimating the unknown filter by a least-squares fit of the reference sensor signal to the primary sensor signal. This method is commonly implemented with the well-known Normalized Least-Mean-Square (NLMS) gradient descent algorithm [2]. Assuming a filter with \( L \) taps and denoting its estimated transfer function at instant \( n \) as \( \hat{h}_n = [h_n(0) \cdots h_n(L-1)] \), the NLMS update equation is

\[
\hat{h}_{n+1} = \hat{h}_n + \mu \frac{(s_1(n) - \hat{h}_n s_2(n)) s_2^*(n)}{\sigma_n}
\]

where the operator * denotes transposition, where \( s_2^*(n) = [s_2(n) \cdots s_2(n - L + 1)] \) is the reference signal input to the filter at time \( n \), where the empirical parameter \( \mu \) controls the convergence speed and where the scalar \( \sigma_n \) is an estimate of the power of the reference signal: \( \sigma_{n+1} = \alpha |s_2(n)|^2 + (1 - \alpha) \sigma_n \) with \( \alpha \) empirically tuned (\( \sigma_1 = (\sum_{l=1}^{L} |s_2(l)|^2)/L \)). In our experiments, the algorithm is used in a batch mode and proceeds as follows.

**Noise canceling algorithm:** for each sentence,
- estimate a filter \( \hat{h}_N \) by iterating equation 1 over all \( N \) samples in the sentence,
- compute the desired signal at time \( n \) as:

\[
d(n) = s_1(n) - \hat{h}_N s_2(n)
\]

3. 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. 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:

\[
P_y(f) = P_x(f) - P_n(f)
\]

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]:

\[
y(t) = x(t) - r(y(t), n(t))
\]

with \( r \) a non linear function. Assuming MFCC vectors computed with a bank of Mel-filters followed by a Discrete Cosine Transform:

\[
r(y(t), n(t)) = DCT \log(1 + e^{DCT^{-1} (n(t) - y(t))})
\]

where \( DCT \) and \( DCT^{-1} \) refer respectively to the Discrete Cosine Transform and to its inverse. For lack of knowing the cepstra \( y(t) \) of the clean speech, the function \( r \) is approximated with its expected value over \( y \), given \( n(t) \) and \( x(t) \):

\[
r(n(t)) = E_y[r(y(t), n(t)) \mid x(t), n(t)]
\]

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 \( C_n = \{ c_i \}_{i=1}^{n_c} \) of \( n_c \) cepstral vectors describing the acoustic space of the clean speech, the noise correction term is computed as:

\[
r(n(t)) = \sum_{i=1}^{n_c} p(c_i \mid x(t), n(t)) \cdot r(c_i, n(t))
\]

Assuming the Gaussian distribution \( \mathcal{N}(y(t) \mid \mu_i, \sigma_i^2) \) 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 \( \mathcal{N}(x(t) \mid \mu_i + r(\mu_i, n(t)), \sigma_i^2) \). The posterior probability of the codeword \( c_i \), given \( x(t) \) and \( n(t) \), is thus computed as:

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

where \( \pi_i \) denotes the a priori probability of the codeword \( c_i \). An estimate of the clean speech \( \hat{y}(t) \) is computed as:

\[
\hat{y}(t) = x(t) - \bar{r}(n(t))
\]
Whereas in standard CDCN, the cepstra of the noise are estimated via an EM algorithm, we propose with MCDCN to compute it from the waveform of the reference signal, i.e. from the signal denoted as $s_2$ in section 2. The cepstra of the noisy speech are computed from the primary signal $s_1$.

**MCDCN algorithm:** at each frame,
- compute the posteriors with equation 7,
- compute the compensation term with equation 6,
- compute the compensated cepstra with equation 8.

### 4. Speech recognition experiments

#### 4.1. Evaluation data

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. All the data were recorded at 22kHz and downsamped to 11kHz. The speech corrupted by the sound emitted by the car speakers was recorded with an AKG Q400 microphone located on the visor. The signals at either the radio output or at the left and right outputs of the CD player were captured in separate channels.

In the experiments presented here, the interfering signal comes from either a mono source (radio) or a stereo source (CD player). 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. The aligned left and right waveforms are then summed up resulting in a single reference signal. The noisy speech waveform and the reference waveform are roughly aligned by detecting the maximum of their cross-correlation function for shifts of up to 90ms. They are scaled by estimating the mean ratio 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).

#### 4.2. Separation protocols

After the preliminary step of alignment and scaling described in section 4.1, the waveforms $s_1$ of the noisy speech and the waveform $s_2$ of the music are processed using either:
- Noise Canceling (NC): the waveform of the clean speech is reconstructed as explained in section 2 with equation 2 assuming a filter with 257 taps\(^1\).

---

\(^1\) with waveforms sampled at 11kHz, a filter with 257 taps allows for delays of up to $\frac{257}{11} \approx 23$ ms, i.e. distances of more than 7 meters between the source of music and the microphone.

- MCDCN (MCDCN): compensated cepstra are computed with equation 8 where the cepstra of the noisy speech and the noise are computed respectively from the primary signal $s_1$ and from the reference signal $s_2$.
- Noise Filtering followed by MCDCN (NF+MCDCN): compensated cepstra are computed with equation 8 where the cepstra of the noisy speech and the noise are computed respectively from the primary signal $s_1$ and from the filtered reference signal $h_N \ast s_2$.

The parameters $\mu$ and $\alpha$ in the NLMS algorithm are empirically chosen so as to maximize the overall speech recognition accuracy. MCDCN is applied with codebooks 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 are assigned equal priors.

#### 4.3. Recognition system

Speech recognition is performed with a reduced-size system especially designed for portable devices or automotive applications [4]. 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 dimensional cepstra (12 MFCC + the energy + delta and delta-delta coefficients) from 15ms frames. However equation 5 is computed using the 24 dimensional MFCC.

#### 4.4. Results

Table 1 shows the baseline average word error rates (WER) obtained by decoding the noisy speech without any compensation and with (NC), (MCDCN) and (NF+MCDCN) using codebooks of size ranging from 1 to 256 codewords. (NC) provides relative WER reductions over the baseline of about 75%, 50% and 40% respectively with the 60, 70 and 80dB interferences. (MCDCN) provides WER that are 3 to 5 times less than with (NC). It achieves its best performance with codebooks of 16 codewords with relative WER reductions over the baseline of about 90%, 90% and 85% with the 60, 70 and 80dB interferences. The combination (NF+MCDCN) provides additional gains: 30% to 40% relative WER reductions over (MCDCN) and as much as 80% to 90% over (NC) depending on the noise level. (NC) and (NF+MCDCN) use the same filtered noise estimate : they differ in that the filtered noise is directly subtracted from the waveform of the noisy speech with (NC), whereas it is used to compute a compensation term.
in the cepstral domain with (NF+MCDCN). Our interpretation is that (NF+MCDCN) compensates for the imprecision on the filter estimate, as well as for the imprecision of the additive noise model, by taking advantage of our a priori knowledge of the clean speech distribution in the cepstral domain. This mechanism allows to obtain substantial recognition accuracy gains with MCDCN in the cepstral domain. This mechanism allows to obtain substantial recognition accuracy gains with MCDCN alone, merely using a single tap delay filter to account for the channel difference between the primary and reference signals. As the (NF+MCDCN) results show, compensating more accurately for the channel difference further improves the recognition accuracy, however most of the gain observed over the baseline with (NF+MCDCN) is brought about by MCDCN alone.

Table 2 shows the breakdown of the average WER for each kind of interfering signal before any compensation and after compensation with the best performing combination of (NF+MCDCN). Not surprisingly, the most confusing interference for the speech recognition system is by far the competing speech from the speaker of the radio news talk, and then the DJ kind of music (note that the DJ tracks used in the experiment consisted of mainly rap music, i.e. again something close to competing speech). But even though in that case both the desired and corrupting signals somehow follow the same clean speech statistical distribution, MCDCN provides as much WER reductions as in the case of jazz or classical music.

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. Unlike adaptive filtering, MCDCN does not focus on estimating the channel difference between the unwanted and desired signals. Instead, it takes advantage of our a priori knowledge of the clean speech distribution. In our speech recognition experiments with real noisy speech, the WER obtained with MCDCN was 3 to 5 times less than with the NLMS noise canceling technique. A combination of the two techniques provided the best performance.

6. References


