MODEL-SPACE COMPENSATION OF MICROPHONE AND NOISE FOR SPEAKER-INDEPENDENT SPEECH RECOGNITION

Yifan Gong

Speech Technologies Laboratory, DSP Solutions R&D Center
Texas Instruments Inc., Dallas, Texas, USA
Email: Yifan.Gong@ti.com

ABSTRACT

Ambient noise (additive distortion) and microphone changes (convolutive distortion) are two sources of distortion that may severely degrade speech recognition performance in real operation environments. Simultaneously modeling the two distortion sources has been a great challenge for robust speech recognition. A method, called JAC, that Jointly compensates both Additive and Convolutive distortions is presented. It uses two log-spectral domain components in speech acoustic models to represent additive and convolutive distortions. The method adapts HMM mean vectors with a noise estimate and a channel estimate. The noise estimate is calculated from the pre-utterance pause and the channel estimate is calculated using an EM algorithm from speech utterances produced in the distortion environment. Evaluated on a noisy speech database recorded in-vehicle with a hands-free distant microphone in several driving conditions, the algorithm reduces recognition word error rate in typical operation conditions by an order of magnitude.

1. INTRODUCTION

A speech recognizer trained with office environment speech data and operating in a different environment may fail due to at least two distortion sources [1]: background noise and microphone changes. Handling simultaneously the two is critical to the performance.

Retraining the recognizer's acoustic model using large amount of training data collected under conditions as close as possible to the testing data could reduce the recognition failure. There are several problems associated with such approach, however. The first is flat distribution and loss of discrimination due to averaging over all conditions. The second is the lack of ability to cover new types of noises and unknown performance for unseen noises. Finally, it is impossible to separate the variabilities at the phonetic level and at the acoustic level.

Model adaptation approach by MAP (maximum a posteriori) [2] estimation or MLLR (maximum likelihood linear regression) [3] transforms HMM parameters to match the distortion factors. It does not model explicitly channel and background noise, but approximates their effect by piecewise linearity. When given enough data, it is effective for both sources. However, typically, the approach requires additional training data collected in the noisy environments, it is optimal only under the type and the intensity of the distortion of the training data, it may introduce dependence to the speaker who provides the training data, and finally, it is useless when the noise changes from utterance to utterance. Therefore, in most applications, such adaptation is not desirable.

Some feature space (front-end) techniques developed in the past provide simple solutions. CMN (cepstral mean normalization) [4] removes utterance mean which is related to channel distortion. SS (spectral subtraction) [5] reduces background noise. Methods such as the ETSI advanced DSR front-end [6] handles both channel distortion and background noise. These techniques do not require any noisy training data. However, to be effective in noise reduction, they typically require an accurate point estimate of the noise spectrum.

In the model space, a convolutive (e.g. channel) component and an additive (e.g. noise) component can be introduced to model the two distortion sources [7, 8, 9, 10, 11, 12]. The effect of the two distortion sources introduces in the log spectral domain non-linear parameter changes, which can be approximated by linear equations [13, 14].

We describe a new framework that handles simultaneously both noise and channel distortions for speaker-independent speech recognition robust to a wide variety of noises and channel distortions.

2. COMPENSATION MODEL

We first establish the relationship between distorted speech and distortion factors.

A speech signal $x(n)$ can only be observed in a given acoustic environment. An acoustic environment can be modeled by a background noise $b'(n)$ and a distortion channel

$h(n)$. For typical mobile speech recognition, $b'(n)$ consists of noise from office, vehicle engine and road noises, and $h(n)$ consists of microphone type and relative position to the speaker. Let $y(n)$ be the speech observed in the environment involving $b'(n)$ and $h(n)$: $y(n) = (x(n)+b'(n))*h(n)$. In typical speech recognition applications, $b'(n)$ cannot be measured directly. What is available is $b'(n) * h(n)$. Let $b(n) = b'(n) * h(n)$, our model of distorted speech becomes:

$$y(n) = x(n) * h(n) + b(n) \tag{1}$$

Or, applying the DFT to both sides of Eq-1:

$$\mathbf{Y}(k) = \mathbf{X}(k)\mathbf{H}(k) + \mathbf{B}(k) \tag{2}$$

In speech recognition systems, typically the parameters are represented in a logarithmic scale. Expressing the above quantities in logarithmic scale, we have:

$$\mathbf{Y}^l \triangleq \mathbf{g}(\mathbf{X}^l, \mathbf{H}^l, \mathbf{B}^l) \tag{3}$$

where

$$\mathbf{g}(\mathbf{X}^l, \mathbf{H}^l, \mathbf{B}^l)(k) = \log(\exp(\mathbf{X}^l(k) + \mathbf{H}^l(k)) + \exp(\mathbf{B}^l(k))).$$

In this paper, we are interested in the distribution of $\mathbf{Y}^l$, given that $\mathbf{X}^l$ is generated by an HMM process and that $\mathbf{H}^l$ and $\mathbf{B}^l$ are two unknown constants.

Assuming the log-normal distribution [15] and ignoring the variance, we have

$$\mathrm{E}\{\mathbf{Y}^l\} \triangleq \hat{\mathbf{m}}^l = \mathbf{g}(\mathbf{m}^l, \mathbf{H}^l, \mathbf{B}^l) \tag{4}$$

where $\mathbf{m}^l$ is the original Gaussian mean vector, and $\hat{\mathbf{m}}^l$ is the Gaussian mean vector compensated for the distortions caused by channel and environment noise. Eq-4 says that the mean vector of a compensated model can be determined if $\mathbf{m}^l$ and an estimate of channel and noise is available.

In the next sections, the solution of the channel estimate is outlined. Detailed description can be found in [16].

2.1. Estimation of noise component

From Eq-1, with $x(n) = 0$, we have: $y(n) = b(n)$ which means that the filtered noise $b(n)$ can be observed during a speech pause. Given that $y(n)$ of the window t is typically represented in MFCC domain as $\mathbf{y}_t$, we can calculate an estimate of noise in the log domain $\mathbf{B}^l$ as the average of P noise frames in the log domain:

$$\mathbf{B}^l = \frac{1}{P}\sum_{t=0}^{P} \mathcal{DFT}(\mathbf{y}_t) \tag{5}$$

2.2. Estimation of channel component

2.2.1. Channel Equation

Our goal is to derive the HMMs of $\mathbf{Y}$ under both additive noise and convolutive distortions. The key problem is to obtain an estimate of the channel $\mathbf{H}^l$. We assume that some speech data recorded in the noisy environment is available, and that the starting HMM models for $\mathbf{X}$ are trained on clean speech in the MFCC feature space.

The speech model is continuous density Gaussian mixture HMMs. In the following, only the mean vectors of the original model space will be modified.

Let R be the number of utterances available for training a hidden Markov model. Let T_r be the number of frames in utterance r. Let Ω_s be the set of states of an HMM. Let Ω_m be the set of mixing components of a state. Let θ_t^r and ξ_t^r be the random variables denoting the index of the state and mixing component, respectively, at time t of the utterance r. According to EM algorithm [17], the auxiliary function of interest can be written as:

$$\mathcal{Q}_b(\lambda|\bar{\lambda}) = \sum_{r=1}^{R}\sum_{t=1}^{T_r}\sum_{j\in\Omega_s}\sum_{k\in\Omega_m} p(\theta_t^r = j, \xi_t^r = k|\mathbf{O}, \bar{\lambda}) \cdot \log p(\mathbf{o}_t^r|\theta_t^r = j, \xi_t^r = k, \lambda) \tag{6}$$

where

$$p(\mathbf{o}_t^r|\theta_t^r = j, \xi_t^r = k, \lambda) \triangleq b_{j,k}(\mathbf{o}_t^r) \sim \mathrm{N}(\mathbf{o}_t^r; \hat{\mathbf{m}}_{j,k}, \mathbf{v}_{j,k})$$

in which $\mathrm{N}(x; \mu, \sigma)$ is a Gaussian distribution with mean vector μ and covariance matrix σ, and

$$p(\theta_t^r = j, \xi_t^r = k|\mathbf{O}, \bar{\lambda}) \triangleq \gamma_t^r(j,k) \tag{7}$$

$\mathbf{o}_t^r$ is the observed speech frame of the utterance r at time t. $\hat{\mathbf{m}}_{j,k}$ is the mean vector of mixing component k of the state j which, according to Eq-4, is a function of $\mathbf{H}$.

To maximize $\mathcal{Q}_b(\lambda|\bar{\lambda})$, differentiating Eq-6 with respect to $\hat{\mathbf{m}}_{j,k}$ and equating to zero, we obtain an equation for each state j and mixing component k. Summing up all these equations, we have

$$\sum_{j\in\Omega_s}\sum_{k\in\Omega_m}\sum_{r=1}^{R}\sum_{t=1}^{T^r} \gamma_t^r(j,k)\mathbf{v}_{j,k}^{-1}\{\hat{\mathbf{m}}_{j,k} - \mathbf{o}_t^r\} = 0 \tag{8}$$

In Eq-8, each $\hat{\mathbf{m}}_{j,k}$ is a function of the channel estimate $\mathbf{H}$. Directly solving this equation for $\mathbf{H}$ involves multi-dimensional non-linear equations and computational heavy transformations between cepstrum and log-spectrum domains. A simplified solution is developed here. Applying DFT, ignoring the variance for simplicity, and using Eq-4, we obtain:

$$\mathbf{u}(\mathbf{H}, \mathbf{B}) = \sum_{j\in\Omega_s}\sum_{k\in\Omega_m}\sum_{r=1}^{R}\sum_{t=1}^{T^r} \gamma_t^r(j,k) \cdot \left\{\mathbf{g}(\mathbf{m}_{j,k}^l, \mathbf{H}^l, \mathbf{B}^l) - \mathcal{DFT}(\mathbf{o}_t^r)\right\} = 0 \tag{9}$$

Eq-9 is called channel equation.

2.2.2. Solving channel equation

Several algorithms can be used to find a solution $\mathbf{H}$ for $\mathbf{u}(\mathbf{H}, \mathbf{B}) = 0$. We use Newton's method, as it has the property of convergence at square rate. Such speed should be in general faster than gradient method. As we are interested in on-line estimation of the parameters, the convergence speed is critical. The method is iterative, which gives a new estimate $\mathbf{H}^l_{[i+1]}$, at iteration $i+1$, of $\mathbf{H}^l$ using:

$$\mathbf{H}^l_{[i+1]} = \mathbf{H}^l_{[i]} - \frac{\mathbf{u}(\mathbf{H}^l_{[i]}, \mathbf{B})}{\mathbf{u}'(\mathbf{H}^l_{[i]}, \mathbf{B})} \tag{10}$$

where $\mathbf{u}'(\mathbf{H}^l, \mathbf{B})$ is the derivative of $\mathbf{u}(\mathbf{H}^l, \mathbf{B})$ with respect to channel $\mathbf{H}^l$. As initial condition for Eq-10, we can set

$$\mathbf{H}^l_{[0]} = \mathbf{0} \tag{11}$$

2.2.3. Compensation for time derivatives

The distortion caused by channel and noise also changes the distribution of dynamic (e.g. time derivative of) MFCC coefficients, in addition to that of static coefficients described so far. Under the present framework, the time derivatives of MFCC can be also compensated.

According to definition, the compensated time derivative of cepstral coefficients $\dot{\mathbf{Y}}^c$ is the time derivative of compensated cepstral coefficients $\mathbf{Y}^c$. It can be shown [16] that both first and second order time derivatives are respectively a function of

$$\eta(k) = \exp(\mathbf{H}^l(k)) \frac{\exp(\mathbf{X}^l(k))}{\exp(\mathbf{B}^l(k))} \tag{12}$$

We recognize that, by definition, $\frac{\exp(\mathbf{X}^l(k))}{\exp(\mathbf{B}^l(k))}$ is the signal-to-noise ratio in linear scale at the the frequency bin k. Consequently, $\eta(k)$ is called the generalized SNR in linear scale at the frequency bin k. Due to space limitation, in this paper the solution for the time derivatives wil not be further elaborated.

3. EXPERIMENTAL RESULTS

3.1. Database and speech models

The database is recorded in-vehicle, using an AKG M2 hands-free distant talking microphone, in four recording sessions: parked-trn (car parked, engine off), parked (car parked, engine off), city (car driven on a stop and go basis), and highway (car driven on highway).

In each session, 20 speakers (10 male) read 40 sentences each, giving 800 utterances. Each sentence is either a 10, 7 or 4 digit sequence, with equal probabilities. There are over 16,000 digit tokens in this test set. The database is sampled at 8kHz, with frame rate of 20ms. From the speech, MFCC of order 10 is derived.

HMMs used in all experiments are trained on TIDIGITS clean speech data. Evaluated on TIGIDIT test set, the recognizer gives 0.36% word error rate.

For the testing hands-free database, the microphone is remote mounted and band-limited, as compared to a high quality microphone used to collect TIDIGITS. Also, there is a substantial amount of background noise due to car environment, with SNR reaching 0dB for the highway driving condition. Compared to the speech database used to train the above HMM models, the database presents severe mismatch to the models trained on TIDIGITS, both in channel and in noise background. It is therefore very challenging to see the performance of different compensation approaches on this database.

In the experiments, the bias is reestimated after recognizing each test utterance. At the beginning of the recognition of all speakers, the bias is set to zero. This scheme makes it possible for the parameter estimation procedure to exploit the information about the channel acquired from recognizing all previous speakers in order to better determine the channel for the current speaker.

3.2. Recognition results

The new algorithm is referred to as JAC (joint compensation of additive noise and convolutive distortion). Table-1 summarizes the recognition performance for BASELINE (no compensation applied), CMN (cepstral mean normalization), PMC (parallel model combination where the noise is represented by a single state HMM), and JAC (on static MFCC coefficients only). Table-1 shows that: /1/ Compared to noise-free recognition WER (0.36%), without any compensation (BASELINE) the recognition performance degrades severely. /2/ CMN effectively reduces the WER for parked data, but is not effective for driving conditions where additive noise becomes dominant. /3/ PMC substantially reduces the WER for driving conditions, but gives poor results for parked data where microphone mismatch is dominant.

	PARKED	CITY	HIGHWAY
BASELINE	1.38	30.3	73.2
CMN	0.59	18.8	51.7
PMC	1.74	6.29	17.0
JAC	0.47	1.84	7.78

Table 1. word error rate (%) as function of driving conditions and compensation methods

/4/ JAC gives substantially lower WER than non-JAC methods.

4. CONCLUSION

The distribution of speech signals observed in a mobile environment is distorted by channel and background noises, compared to the speech acoustic models trained in a clean environment. Handling simultaneously the two sources of distortion is critical to maintaining the recognition performance in typical situations where the recognizer is deployed.

Automatically adjusting speech model parameters according to these distortions is one way of modeling the two sources. A recognition method is developed which identifies two log-domain components from incoming speech signal: one for the channel or microphone distortion (convolutive), and the other for the background noise (additive). An EM algorithm based procedure is derived to estimate the convolutive component iteratively. The method compensates for the utterance-specific distortions by modifying mean vectors of the HMMs of the recognizer for every incoming utterance.

Experimental results show that, although very simple, the method is extremely efficient in improving speaker-independent recognition performance in a real application task. The method substantially reduced the word error rates obtained using either CMN or PMC, and achieved an overall 86% average word error rate reduction compared to baseline performance. The method makes it possible to obtain high performance for speaker-independent recognition in changing noisy environments without explicitly collecting any noisy speech for training.

5. REFERENCES

[1] Y. Gong, "Speech recognition in noisy environments: A survey," *Speech Communication*, vol. 16, no. 3, pp. 261–291, April 1995.

[2] J. L. Gauvain and C. H. Lee, "Maximum a posteriori estimation for multivariate gaussian observations of Markov chains," *IEEE Trans. on Speech and Audio Processing*, vol. 2, no. 2, pp. 291–298, April 1994.

[3] C. J. Leggetter and P. C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density HMMs," *Computer, Speech and Language*, vol. 9, no. 2, pp. 171–185, 1995.

[4] B. Atal, "Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identifition and verification," *J. of Acoust. Soc. America*, vol. 55, pp. 1304–1312, 1974.

[5] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. on Acoust., Speech and Signal Processing*, vol. ASSP-27, no. 2, pp. 113–120, April 1979.

[6] D. Macho, L. Mauuary, B. Noe, Y. M. Cheng, D. Ealey, D. Jouvet, H. Kelleher, D. Pearce, and F. Saadoun, "Evaluation of a noise-robust DSR front-end on AURORA databases," in *Proc. Int. Conf. on Spoken Language Processing*, Colorado, USA, September 2002, pp. 17–20.

[7] M. Afify, Y. Gong, and J.-P. Haton, "A general joint additive and convolutive bias compensation approach applied to noisy lombard speech recognition," *IEEE Trans. on Speech and Audio Processing*, vol. 6, no. 6, pp. 524–538, November 1998.

[8] Y. Zhao, "Frequency-domain maximum likelihood estimation for automatic speech recognition in additive and convolutive noises," *IEEE Trans. on Speech and Audio Processing*, vol. 8, no. 3, pp. 255–266, May 2000.

[9] J. L. Gauvain, L. Lamel, M. Adda-Decker, and D. Matrouf, "Developments in continuous speech dictation using the ARPA NAB news task," in *Proc. of IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, Detroit, 1996, pp. 73–76.

[10] Y. Minami and S. Furui, "A maximum likelihood procedure for a universal adaptation method based on HMM composition," in *Proc. of IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, Detroit, 1995, pp. 129–132.

[11] M. J. F. Gales, "PMC for speech recognition in additive and convolutional noise," Tech. Rep. TR-154, CUED/F-INFENG, December 1993.

[12] Y. Gong, "A robust continuous speech recognition system for mobile information devices (invited paper)," in *Proc. of International Workshop on Hands-Free Speech Communication*, Kyoto, Japan, April 2001.

[13] S. Sagayama, Y. Yamaguchi, and S. Takahashi, "Jacobian adaptation of noisy speech models," in *Proceedings of IEEE Automatic Speech Recognition Workshop*, Santa Barbara, CA, USA, DEC 1997, pp. 396–403, IEEE Signal Processing Society.

[14] N. S. Kim, "Statistical linear approximation for environment compensation," *IEEE Signal Processing Letters*, vol. 5, no. 1, pp. 8–10, Jan. 1998.

[15] M. J. F. Gales and S. Young, "An improved approach to the hidden Markov model decomposition of speech and noise," in *Proc. of IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, U.S.A., April 1992, vol. I, pp. 233–236.

[16] Y. Gong, "A method of joint compensation of additive and convolutive distortions for speaker-independent speech recognition," *IEEE Trans. on Speech and Audio Processing*, (submitted for publication) 2002.

[17] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society*, vol. 39, no. 1, pp. 1–38, 1977.