STATISTICAL ESTIMATION FOR HANDS-FREE SPEECH RECOGNITION

Yunxin Zhao

Department of Computer Engineering and Computer Science
University of Missouri, Columbia, MO 65211
zhao@cecs.missouri.edu

ABSTRACT

This work reports a cross-condition speech recognition experiment between TIMIT and FFMTIMIT. The TIMIT data were collected by a head-mounted close-talking microphone and was used as training speech for a speaker-independent continuous speech recognition system, and the FFMTIMIT data were collected by a far-field microphone and were used as test speech. The condition mismatch between the two databases consists of both channel distortion and additive noise. A frequency-domain EM algorithm was employed for online identification of channel and noise parameters and for estimation of clean speech features from FFMTIMIT speech. The statistical estimation algorithm has led to a significant performance improvement for the cross-condition recognition task.

1. INTRODUCTION

A large number of speech databases that are widely used in speech research community were collected by head-mounted closed-talking microphones. These speech data are of high quality, well processed and documented, and hence provide a valuable resource for training statistical models of speech. On the other hand, tether-free speech acquisition is preferred or needed in many applications of spoken language technology. It is therefore of practical and theoretical interests to address the problem of training models using speech collected by closed-talking microphones and test recognition on speech acquired by free-field microphones. The technological challenge is well known to lie in acoustic environment mismatch as described by the system of Fig. 1.

In Fig. 1, $x$ denotes clean speech acquired by a close-talking-microphone, $y$ denotes noisy speech acquired by a free-field microphone. The two types of microphones in general have different frequency responses. The increased length of acoustic path between a talker and a free-field microphone introduces into speech a higher level of background noise $v$ and a frequency-shaping distortion. Therefore, $x$ is transformed into $y$ by the combined effects of channel distortion (convolutive noise) and additive noise, and the mismatch between $x$ and $y$ causes degradation of recognition performance.

Several techniques have been proposed for robust speech recognition in additive and convolutive noises. The statistical-estimation-based approaches consist of SNR-dependent compensation of acoustic models in log spectral domain [1], iterative bias estimation in both spectral and cepstral domains [2,3], vector Taylor series (VTS) approximation for model compensation in log spectral domain [4], channel estimation in linear or cepstral domains with parallel model compensation (PMC) for noise [5,6], PMC in log spectral or cepstral domain [7], frequency-domain EM-based joint estimation of channel and noise [8], hidden Markov modeling (HMM) of noise and channel for dynamic composition of speech models in time-varying environment [9].

In this work, an experimental investigation is made on using the statistical estimation technique of [8] to resolve an environment mismatch problem between two databases, i.e., TIMIT and FFMTIMIT. Both databases are released by the Linguistic Data Consortium (LDC) of University of Pennsylvania of USA [http://www.ldc.upenn.edu]. Specifically, the TIMIT database [10] was collected by a head-mounted closed-talking microphone and the FFMTIMIT database [11] was collected by a far-field microphone. The FFMTIMIT data have a significant amount of low-frequency noise and were not identically processed as the TIMIT data when released on the CDs. The TIMIT data were used to train HMM phone models of a speaker-independent continuous speech recognition system, and the FFMTIMIT data were used to evaluate system performance. A direct recognition on the FFMTIMIT data by the TIMIT-trained system yielded a very poor result. The EM algorithm is shown to have significantly improved recognition accuracy on FFMTIMIT data using the TIMIT-trained system. This experimental study extends the work of [8] into a real-world application of hands-free speech recog-
tion.

The rest part of the paper is organized as follows. The background of the frequency domain EM algorithm is provided in Section 2. The experiments on the FFM/TIMIT data set is described in Section 3. A conclusion is made in Section 4.

2. FREQUENCY-DOMAIN EM ALGORITHM

The speech degradation mechanism shown in Fig. 1 can be described in the discrete Fourier transform (DFT) domain as

\[ Y_n(\omega) = \Theta(\omega)X_n(\omega) + V_n(\omega) \]

with \( n \) denoting the short-time analysis frames. \( Y_n(\omega), X_n(\omega), V_n(\omega) \) the short time DFTs of degraded speech, clean speech, and noise, and \( \Theta(\omega) \) the DFT of channel’s finite impulse response \( \theta \).

For a size-\( N \) DFT, the probability density function (pdf) of clean speech is assumed to be a mixture of Gaussian densities

\[ f_X(X_n(\omega); \Lambda X) = \sum_{i=1}^{M} \alpha_i \prod_{i=0}^{N-1} \mathcal{N}(X_n(\omega_l); 0, \Phi_{X X,i}(\omega_l)) \]

where \( \alpha_i \)'s are mixture weights, \( 0 \)'s denote zero-mean which is inherited from the zero-mean property of speech waveform, and \( \Phi_{X X,i}(\omega_l) \)'s are class-conditional spectral variances that characterize the spectral distributions of speech. The model parameters \( \Lambda X \) can be estimated from clean training speech. Noise is assumed to be from a single Gaussian source with the spectral variances \( \Phi_{V V}(\omega) \). The pdf of degraded speech is the same as clean speech except that the spectral variances are changed to \( \Phi_{Y V,Y}(\omega) = |\Theta(\omega)|^2 \Phi_{X X,i}(\omega) + \Phi_{V V}(\omega) \). The conditional pdf of clean speech given degraded speech can be derived as

\[ f_{X|Y}(X_n(\omega)|Y_n(\omega); \lambda) = \sum_{i=1}^{M} \alpha_i \prod_{i=0}^{N-1} \mathcal{N}(X_n(\omega_l); \mu_{X,Y,i}(\omega_l), \Phi_{X X,i}(\omega_l)) \]

where the unknown environment parameters are denoted by \( \lambda = (\theta, \Phi_{v v}(\omega)) \).

The identification problem is defined as estimation of \( \lambda \) from a sequence of degraded speech vectors

\[ Y_n^T(\omega) = (Y_1(\omega), Y_2(\omega), \ldots, Y_T(\omega)) \]

A sequence of complete vectors is denoted as \( Z_n^T(\omega) \), with the definition of \( Z_n(\omega) = (Y_n(\omega), X_n(\omega)) \). Maximum likelihood estimation of \( \lambda \) is achieved by the EM procedure. At the k-th step, the auxiliary function

\[ Q(\lambda, \lambda^{(k)}) = E \left[ \log f_Z(Z_n^T(\omega); \lambda) | X_n^T(\omega), \lambda^{(k)} \right] \]

is maximized to obtain \( \lambda^{(k+1)} \), for \( k = 0, 1, \ldots \) until convergence.

Each iteration of EM consists of an expectation step and a maximization step. In the k-th iteration, the expectation consists of computing the conditional statistics

\[ \Psi_{X X|Y_n}(\omega) = \sum_{i=1}^{M} \alpha_i \Phi_{X X,i}(\omega) \]

\[ C_{X X|Y_n}(\omega) = E \left[ |X_n(\omega)|^2 | Y_n(\omega); \lambda^{(k)} \right] \]

\[ = \sum_{i=1}^{M} \alpha_i \Phi_{X X,i}(\omega) + \mu_{X X,i}(\omega)^2 \]

where \( \Psi_{X X|Y_n}(\omega) \) is the average posterior spectral variance, and \( C_{X X|Y_n}(\omega) \) is the posterior estimate of power spectrum. In the maximization step, the new parameter estimate \( \lambda^{(k+1)} \) is calculated by solving the root of

\[ \frac{\partial Q(\lambda, \lambda^{(k)})}{\partial \lambda} = 0 \]

which is a function of the conditional statistics, the previous parameter estimate \( \lambda^{(k)} \), and the input spectral mean \( \frac{1}{n} \sum_{n=1}^{N} |Y_n(\omega)|^2 \). Due to the choice of DFT domain, closed-form solutions of \( \lambda^{(k+1)} \) is obtained in each iteration. At the convergence of EM, the posterior short-time power spectral estimates can be used to estimate speech features for recognition.

Let \( f(\|X_n(\omega)\|^2) \) be the function of computing a feature vector \( s_n \) from a spectrum \(|X_n(\omega)|^2\). The minimum mean-squared-error (MMSE) estimator of the feature vector is

\[ s_n = \sum_{i=1}^{M} \alpha_i Y_n E[f(\|X_n(\omega)\|^2) | Y_n(\omega), I_n = i] \]

By first-order Taylor series expansion, the feature estimate is approximated as

\[ s_n \approx \sum_{i=1}^{M} \alpha_i Y_n f(\Phi_{X X,i}(\omega) + |\mu_{X X,i}(\omega)\|^2) \]

Dynamic speech features can be computed from the estimated instantaneous feature sequence \((\hat{s}_1, \hat{s}_2, \ldots, \hat{s}_T)\) by temporal regression.

For further details of the estimation algorithm, please refer [8].

3. EXPERIMENTS

The TIMIT and FFMTIMIT data were simultaneously recorded and were designated by LDC as the primary and secondary waveforms, respectively. The TIMIT data were recorded using a close-talking noise-cancelling head-mounted Sennheiser microphone, and the FFMTIMIT data
were recorded using a Breul & Kjaer 1/2" free-field microphone. Unfortunately, no documentation can be found on the distance between speaker and the free-field microphone, and the distances were recalled to be in the range of 25 cm to 35 cm [12].

Compared with the TIMIT data, many FFMTIMIT data files are very noisy. The noise was due to an HVAC system and mechanical vibration transmitted through the floor of the double-walled sound booth used in recording. In order to quantify the noise condition in FFMTIMIT, an SNR is computed for each sentence recording file as

$$\text{SNR} = 10 \log_{10} \left( \frac{\hat{E}_y - \hat{E}_n}{\hat{E}_n} \right)$$  \hspace{1cm} (1)

where $\hat{E}_y$ is the noisy speech energy, and $\hat{E}_n$ is the noise energy computed in the same recording file outside the speech end-points. Fig. 2 shows a pair of waveforms of two simultaneously recorded sentences. The SNR of the TIMIT sentence is measured as 39.3 dB, and the SNR of the FFMTIMIT sentence is 2.9 dB.

Figure 2. Two simultaneously recorded sentences.

In Fig. 3, an average noise spectrum was computed from 3000 samples of the noise tail in the FFMTIMIT sentence recording of Fig. 2. The spectrum was computed using a 512-point FFT with a shift of 512 points, where the unit of frequency axis is \( \pi \) which corresponds to the half sampling frequency of 3.4 KHz. The noise energy is seen to be concentrated in the low frequency range.

![Figure 3. An average noise spectrum computed from the FFMTIMIT file of Fig. 2.](image)

In the aspect of channel distortion, the close-talking and free-field microphone have somewhat different frequency responses and only the TIMIT data were processed by a 1581-point high-pass filter.

Both TIMIT and FFMTIMIT data were downsampled from 16 KHz to 6.8 KHz. The TIMIT database was used to train HMM phone models of a speaker-independent continuous speech recognition system [13]. The training data consisted of 717 sentences spoken by 224 male and 101 female speakers. The analysis frame size was 200 samples (29.4 ms) and the shift between adjacent frames was 64 samples (9.4 ms). For each utterance, the speech samples were scaled by square-root of the maximum frame energy of the utterance [8]. Short-time power spectra were computed using 256-point FFTs with zero-padding and without tapering window. The recognition system used the features of perceptually-based linear predictive coding cepstral coefficients, log energy, and first-order temporal regressions of the instantaneous features. The recognition vocabulary was 853 and the task perplexity was 64.

Two test sets were formed, one from the FFMTIMIT, another from the TIMIT, with each set consisting of 177 sentences that were simultaneously recorded with the other set. In the FFMTIMIT test set, the sentence SNRs as measured by Eq.(1) vary significantly. In Fig. 4, the SNR histogram is shown for the FFMTIMIT test set. It is seen that the SNRs cover a broad range of $-10$ dB to 19 dB, with the average SNR of 7.3 dB and a standard deviation of 4.7 dB. However, the SNRs can only be considered as a rough estimates since for each sentence recording file, noise energy inside speech region is unavailable and is likely to fluctuate.
The recognition word accuracy when measured on the TIMIT test set was 86.5%, and that on the FFMTIMIT was 12.6% (BL). By using the EM algorithm in an utterance-by-utterance fashion with 10 iterations in each utterance, the word accuracy increased to 80.7% (EM(10)). In comparison, the method of cepstral bias removal yielded an accuracy of 53.6% (CBR), and spectral subtraction combined with cepstral bias removal (SSCBR) yielded an unexpectedly poor result of 28%. The ineffectiveness of SSCBR was due to the somewhat nonstationary noise characteristics in the FFMTIMIT, where it was observed that in a large number of sentence recordings, the noise spectra measured immediately before speech onset exhibited significant time variations. In order to evaluate the effect of channel distortion on recognition, the EM algorithm was also run without estimating the channel, i.e., the channel was fixed as a unit impulse response. The word accuracy achieved in this case was 66.2% (EN). Comparing the results of EM(10) and EN, the channel effect is significant. These results are summarized in Table 1.

Table 1. Word accuracies (%) on the FFMTIMIT test set using various methods.

<table>
<thead>
<tr>
<th></th>
<th>BL</th>
<th>CBR</th>
<th>SSCBR</th>
<th>EM(10)</th>
<th>EN</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>12.6</td>
<td>53.6</td>
<td>28.0</td>
<td>80.7</td>
<td>66.2</td>
</tr>
</tbody>
</table>

4. DISCUSSION

In the current work, a cross-condition speech recognition experiment between TIMIT and FFMTIMIT is conducted. It is observed that in the same acoustic environment, speech data acquired by a free-field microphone (FFMTIMIT) have significantly lower SNRs than data collected by a close-talking microphone (TIMIT). Due to the acoustic mismatch across the two data sets, a direct recognition of FFMTIMIT speech by a system trained on TIMIT data led to very poor results. In order to reduce the acoustic mismatch between the two data sets, the frequency-domain EM algorithm proposed in [8] was used for estimation of speech features from the FFMTIMIT speech and has led to significant reduction of recognition errors. This result indicates the potential of training hands-free speech recognition system by clean speech models.

ACKNOWLEDGMENT

This work is partially supported by the National Science Foundation under the grant NSF-EIA–9911095.

REFERENCES