Robust Speech Recognition in Distant Environment Based on Speaker Position and Speaking Direction Detection

Longbiao Wang, Naoko Kakutani, Norihide Kitaoka, Seiichi Nakagawa

Department of Information and Computer Sciences, Toyohashi University of Technology
1-1, Hibarigaoka, Tempaku-cho, Toyohashi, Aichi 441-8580, Japan
{xiaowang, naoko, kitaoka, nakagawa}@slp.ics.tut.ac.jp

Abstract

In a practical environment, channel distortion may severely degrade speech recognition performance. In this paper, we propose a robust speech recognition method using real-time Cepstral Mean Normalization (CMN) [1] based on speaker position and speaking direction detection. We first estimate the speaker position in a 3-D space based on the time delay of arrival (TDOA) between distinct microphone pairs. Simultaneously, the system detects the speaking direction of the speaker. We propose a method to estimate the speaking direction based on the radiation characteristics of speech from human’s mouth. And then the system can apply a channel distortion compensation method to the speech using known transmission characteristics of the position and the direction. We measured the transmission characteristics from some grid points in the room and those from human’s speaking positions to various directions a priori, and the system selects the characteristics from them corresponding to the position and the direction. A new real-time CMN based on speaker position estimation was used to reduce the channel distortion. Our experiments show an improvement of 6.5% compared to the baseline system, 10.2% on conventional CMN, 0.8% on real-time position independent CMN.

1. Introduction

In a distant environment, channel distortion may rapidly degrade speech recognition performance. This is mostly caused by the mismatch between the practical environment and training environment. Compensating the input features is a main way to reduce the mismatch. CMN has been used to reduce the channel distortion as a simple yet effective way of normalizing the feature space. CMN provide an error rate reduction under mismatched conditions, and it is very simple to implement too. So many current systems have adopted it. However the system should wait until the end of speech to start the recognition procedure when adopting a conventional CMN. In [1], the CMN was modified to estimate compensation parameters from past few results for real-time recognition. But in a distant environment, the transmission characteristics from different speaker position and speaking direction are very different. This means that the method in [1] cannot compensate the mismatch in the context of hands-free speech recognition. In this paper, we propose a robust speech recognition method using a new real-time CMN based on speaker position and speaking direction detection, which we call real-time position dependent CMN (RTPD CMN). We first estimate the speaker position in a 3-D space based on microphone arrays. Four microphones were arranged in T-shaped on a plane and the sound source position was estimated by TDOA among the microphones estimated using cross-correlation. Our method does not assume that the direction of speaker from the microphones is in parallel each other, so the position can be estimated accurately. Simultaneously, the system detects the speaking direction of the speaker. We propose a method to estimate the speaking direction based on the radiation characteristics of speech from human’s mouth. We set four microphone surrounding the speaker and estimate the direction using ratios of speech powers received by them. We measured the transmission characteristics (real-time position dependent CMN) from some grid points in the room and those from human’s speaking positions to various directions a priori, and system selects the characteristics from them corresponding to the position and the direction. And then the system can apply a channel distortion compensation method to the speech using known transmission characteristics of the position and the direction. The real-time position dependent CMN was used to reduce the channel distortion.

2. Speaker position estimation

It has been shown that it is possible to calculate the speaker position based on the time delay of arrival (TDOA) between distinct microphone pairs. In this paper, TDOA is based on cross-correlation. Conventional sound source estimation method assume that the direction of speaker from the microphones is in parallel each other, so the position can be only approximately estimated even
if the TODA is correct. In this paper, we estimate the speaker position as follows.

It is assumed that N microphones are located at position \((x_i, y_i, z_i)\). And sound source is located at \((x_s, y_s, z_s)\). The distance between the sound source and the \(i\)th microphone is denoted by

\[
D_i = \sqrt{(x_i - x_s)^2 + (y_i - y_s)^2 + (z_i - z_s)^2}.
\]  

(1)

The difference in the distances of microphones \(i\) and \(j\) from the sound source is given by

\[
d_{ij} = D_i - D_j = c\tau_{ij},
\]

where \(c\) is the speed of sound and \(\tau_{ij}\) is the time delay of arrival (TDOA). TDOA can be estimated by cross-correlation. cross-correlation is given by

\[
R_{x_i x_j}(\tau_{ij}) = \frac{1}{T-\tau} \int_{\tau}^{T} x_i(t)x_j(t-\tau_{ij}) \, dt,
\]

(3)

where \(x_i\) and \(x_j\) are the sound signal received by a microphone pair.

And TDOA \(\tau_{ij}\) is given by

\[
\tau_{ij} = \arg \max_{\tau_{ij}} R_{x_i x_j}(\tau_{ij}).
\]

(4)

In order to estimate the speaker position in a 3-D space, four microphones are required. Microphone array is set on a plane as Fig. 1. We can estimate the speaker position by using three microphone pairs \((M1, M2), (M1, M3), (M1, M4)\), respectively. The first microphone \((M1)\) is regarded as the reference and is placed at the origin of the coordinate system. The other three microphones are placed on the plane and have the same distance \(d\) from the first microphone \((M1)\). Because of the symmetry of three microphone pairs, three equations given by (2) with square root can be simply solved as:

\[
x^2 + a_1y^2 + z^2 + b_1y + c_1 = 0
\]

\[
x^2 + y^2 + a_2z^2 + b_2z + c_2 = 0
\]

\[
x^2 + y^2 + a_3z^2 + b_3z + c_3 = 0
\]

where \(a_1, b_1, c_1, a_2, b_2, a_3, b_3\) and \(c_3\) are determined by the distance \(d\) of microphone pair and the TDOA. And the sound source coordinate can be simply obtained by solving the quadratic equations.

3. Speaking direction detection

Speaking direction detection can be used not only to compensate the channel distortion caused by the direction, but also to be used as a modality of interface. In conventional speech recognition system, the system does not begin to work unless the speaker presses the talk switch control unit. It is troublesome for user to do that. To solve this problem, in this paper, speaking direction detection was introduced into a speech dialog system.

The distributions of sound pressure are changed by the speaking angle and frequency [2]. Four microphones were set surrounding the speaker. The utterances toward four microphones \((M1, M2, M3 \text{ and } M4)\) are assumed as \(U_1, U_2, U_3 \text{ and } U_4\), respectively. The system determines the direction using ratios of speech powers received by them.

Training data spoken toward four microphones was recorded a priori. The power of each utterance was calculated from some segments of the utterance. The power ratio of the voiced segments between two microphones was calculated and expressed the power ratio as \(x\) in a logarithmic scale. Assumed that those log power ratios obey the Gaussian distribution \(P(x | U_i) = N(\mu_i, \Sigma_i)(i = 1, 2, 3, 4)\), we can estimate the mean and variance for every direction. In a speaking direction detection stage, we calculate the probability from each frame and the direction with maximum probability is regarded as estimated direction of utterance.

4. Real-time position dependent CMN

4.1. Real-time CMN

A simple and effective way of channel normalization is to subtract the mean of each cepstral coefficient (CMN) which will remove time-invariant distortions introduced by the transmission channel and the recording device. Convolutional noise is considered as additive noise in the cepstral domain, so the noise (transmission characteristics or channel distortion) can be reduced by CMN in the cepstral domain as:

\[
\tilde{c}_t = c_t - \Delta C, \quad (t = 0, ..., T).
\]

(5)

When using conventional CMN, the compensation parameter \(\Delta C\) can be calculated at the end of input speech. This avoids real-time processing of speech recognition. The other problem of conventional CMN is that the number of cepstral coefficients of each utterance are too little.

So conventional CMN cannot adapt the real distant environment well especially when the utterance is short.

We solve the problem under the assumption that the chan-
nel distortion does not change drastically. In our method, the compensation parameter is calculated from utterances recorded a priori. The new compensation parameter is defined by
\[ \Delta C = \bar{C}_{\text{noisy}} - \bar{C}_{\text{clean}}, \] (6)
where \( \bar{C}_{\text{noisy}} \) and \( \bar{C}_{\text{clean}} \) are means of cepstrums of utterances at the position and training clean speech, respectively. Using this method, the compensation parameter can be calculated before recognition of current utterance. Moreover, as the new CMN compensation parameter is estimated by enough number of cepstral coefficients of utterances, so it can compensate the environment distortion better than the conventional CMN. We call this new CMN real-time CMN.

4.2. Incorporate speaker position estimation and speaking direction detection into real-time CMN
CMN, especially real-time CMN, can improve the performance of speech recognition system efficiently. However, in a real distant environment, the transmission characteristics of different speaker position and different speaking direction are very different. So the performance of speech recognition system based on conventional CMN or real-time CMN will be rapidly degraded because of the drastically change of channel distortion.

In this paper, we incorporate speaker position and speaking direction into real-time CMN. We call this method real-time position dependent CMN. We divide the room into some areas and measure those transmission characteristics of each area. We estimate transmission characteristics (the compensation parameter for real-time CMN) using utterances from some grid points in the room and those from human’s speaking positions to various directions parameter from them corresponding to the speaker position and the speaking direction by using speaker position estimation method proposed in Section 2 and speaking direction detection method proposed in Section 3. Our experimental results on real-time CMN combined with speaker position show better than conventional CMN and conventional real-time CMN. And the experiment of real-time CMN combined with speaking direction will be performed in the future.

5. Experimental results

5.1. Speaker position estimation results
The distance \( d \) of microphone pair as shown in Fig. 1 is 20cm. The sound signal received by four microphones are sampled at 48kHz. The experiments of speaker position estimation proposed at Section 2 are performed. The sound source positions are \((0, 120, -50)\), \((60, 103.9, -50)\), \((84.9, 84.9, -50)\), \((103.9, 60, -50)\), \((120, 0, -50)\). Relative precision (RP) of estimated position and real sound source position is defined by
\[ RP = 1 - \frac{\sqrt{(x_r - x)^2 + (y_r - y)^2 + (z_r - z)^2}}{\sqrt{x_r^2 + y_r^2 + z_r^2}}, \] (7)
where \((x_r, y_r, z_r)\) is the real sound source coordinate and \((x, y, z)\) is the estimated sound source coordinate. In our experiments, \( RP = 0.745 \). The result is not very accurate because of the TDOA based on cross-correlation cannot estimate the time delay very accurately. However, in our method we can estimate the speaker position under the error of 50cm under the environment that the distance of sound source is about 200cm far from the microphone array. So the result can be used to determine what area the speaker should be and select the transmission characteristics of this area to compensate the channel distortion. Moreover, more accurate TDOA method such as Crosspower-Spectrum Phase (CSP) etc. can be used to estimate the more accurate time delay and so we will obtain the more accurate speaker position.

5.2. Speaking direction detection results
The sampling frequency was 16kHz. The frame length was 256 points and frame shift was 128 points. In our experiment, 50 frames were used to estimate the speaking direction.
10 speakers (7 males and 3 females) spoke five Japanese vowels or three classes of short sentences about 4 seconds to various directions. For the recorded data spoken by 10 speakers, the utterances spoken by 5 speakers were used as training data and the utterances spoken by the other 5 speakers were used as test data. And the experimental results were showed in Table 1. Experiment showed that most speaking directions could be estimated correctly by our method.

| Table 1: Results of estimated speaking direction |
|-------------|-------------|-------------|-------------|
| U1          | U2          | U3          | U4          |
| 0.998       | 0.001       | 0           | 0           |
| 0.068       | 0.851       | 0.080       | 0           |
| 0           | 0.113       | 0.886       | 0           |
| 0.088       | 0.136       | 0.119       | 0.655       |

5.3. Real-time position dependent CMN results
The transmission characteristics of different speaker position and speaking direction are very different because of the distance between the speaker and microphone array and the reverberation of the room. So we divided the room into 9 \((3 \times 3)\) areas and each area is a 60cm*60cm rectangle. Area 1, area 2 and area3 are near the microphone array and area 7, area8 and area9 are far from microphone array than them. And then we measured the transmission characteristics (that is mean of cepstrums of
utterances recorded a priori) of each area from the center of the area. The system can select the transmission characteristics from them corresponding to speaker position estimated by the method proposed in Section 2.

We evaluated the method on small vocabulary isolated word recognition of 100 words. Here, we assumed that the position was accurately estimated and purely evaluated the real-time CMN. The training set includes 15792 utterances spoken by 158 speakers. The test set utterances are spoken by 5 speakers in a room. Each speaker spoke 200 isolated words. For the utterances of each speaker, 100 utterances were used for estimating the mean of cepstrums of utterances recorded a priori and the other 100 utterances were used for recognition. In our experiment, all speech was emitted from a loudspeaker of the center of each area. The sampling frequency was 12kHz. The frame length was 21.3ms and frame shift is 8ms with 256 points Hamming window. 114 three-state left-to-right syllable based HMMs were used as acoustic models. The feature space was comprised of 10 mel frequency LPC cepstral coefficients. First and second order differentials plus first and second differentials of power component were also included.

The new method is referred to as RTPD CMN (Real-Time Position Dependent CMN). In Table 2, the RTPD CMN is compared with baseline (recognize with no CMN), conventional CMN and RTPI CMN (Real-Time Position Independent CMN), which means the averaged compensation parameters over 9 areas. Both real-time position dependent CMN and real-time position independent improve the performances than the baseline. Real-time position dependent CMN improves all performances of each area except area 3. Table 2 also shows that the more improvement was achieved if the distance between sound source and microphone array was farther. The 9th row in Table 2 is the relative improvement (RI) compared to the baseline. Conventional CMN degraded the recognition rate because of the distant environment and reverberation of the room. Experimental results show that our method relatively improves the recognition rate by 6.5% compared to the baseline system, in other words, the error reduction rate was 51.3%. The RTPD CMN worked better than conventional one and RTPI CMN, because the transmission of the room are very different and so the RTPD CMN could estimate the compensation parameters more accurately than others.

### 6. Conclusion and future work

We proposed a channel distortion compensation method based on real-time position dependent CMN. At first we measured the transmission characteristics from some grid points in the room to estimate the compensation parameters for CMN. Then we estimated the speaker position based on time delay of arrival between microphone pair estimated using cross-correlation. Simultaneously, the speaking direction of speaker could also be detected. Using the speaker position, the system could select the compensation parameter for current utterance. The proposed method could improve the recognition rate better than not only conventional CMN but also real-time position independent CMN.

In our future work, more accurate TDOA method Crosspower-Spectrum Phase (CSP) etc. will be used to estimate the more accurate time delay. The experiment of real-time position independent CMN combined with speaking direction detection will be performed in the future. At last, we will try to measure the impulse response of the room in different areas by using TSP (time-stretched pulse) method. The TSP is a chirp-like signal having a flat over overall power spectrum, that enables a very accurate measurement of the acoustic impulse response. It had been proposed in [4] that training of HMM with filtered speech material by TSP could perform better than training of HMM with clean speech of the same database. Using measured impulse response from each area, we can compensate the channel distortion of each area.

### 7. References

