Multiple-Microphone Robust Speech Recognition Using Decoder-Based Channel Selection

Yasunari Obuchi

Advanced Research Laboratory
Hitachi Ltd., Tokyo, Japan
obuchi@rd.hitachi.co.jp

Abstract

In this paper, we focus on speech recognition using multiple microphones with varying quality. The quality of one channel may be much better than other channels and even the output of standard microphone array techniques such as the delay-and-sum beamformer. Therefore, it is important to find a good indicator to select a channel for recognition. This paper introduces Decoder-Based Channel Selection (DBCS) that gives a criterion to evaluate the quality of each channel by comparing the speech recognition hypotheses made from compensated and uncompensated feature vectors. We evaluate the performance of DBCS using speech data recorded by a PDA-like mockup. DBCS with Delta-Cepstrum Normalization for single channel compensation provides significant improvement compared to the delay-and-sum beamformer. In addition, the concept of DBCS is extended to the delay-and-sum beamformer outputs of various subset of microphones. This extension gives some additional improvement of the speech recognition accuracy.

1. Introduction

It is well known that the performance of automatic speech recognition systems degrades when they are used in noisy environments. There have been huge efforts to solve this problem, which include single-channel feature compensation and microphone array processing. In the single-channel case, input feature vectors are normalized using statistical assumptions for the speech or the noise model. Recently we proposed a novel algorithm called Delta-Cepstrum Normalization (DCN) [1], which is an extension of Histogram Equalization (HEQ) [2] to the cepstral time-derivative domain. It was shown that DCN provides better compensation than Cepstral Mean Normalization (CMN) [3] and HEQ, especially in highly noisy environments.

If the system has more than one microphone, geometric separation of the speech and the noise is possible. A typical approach is the delay-and-sum beamformer [4], in which the speech from the specific direction is enhanced by adding phase-matched signals, while the noises are reduced by averaging phase-unmatched signals. The concept of the delay-and-sum beamformer is based on the assumption that the microphones are homogeneous; the only difference is the geometric position that makes a small difference of the phase of input signals, and all the other conditions are equal. This assumption may hold if the microphones are placed firmly and maintained in a good condition. However, there are some cases where the assumption does not hold. If one holds a PDA that has microphones at each of the four corners and speaks to it, one microphone can be much closer to the speaker's mouth than others. In addition, a finger of the speaker may interfere with a microphone. Similarly, that kind of problem can occur in an automobile, if microphones are placed at various places. In such cases, the input signals have different characteristics, and the quality of the delay-and-sum beamformer output is not always better than that of the best single microphone. However, we have no way of finding out which one is the best.

This paper proposes a new algorithm to select a suitable channel for speech recognition using the output of the speech recognizer. A single-channel feature compensation method is applied to each channel, and both the compensated and uncompensated features are fed into the speech recognizer. The comparison of two outputs gives the estimation of the degree of corruption of the original input. In this paper, DCN is used as the single-channel feature compensation method, but Mean and Variance Normalization (MVN)[5] and HEQ are also investigated.

The remainder of this paper is organized as follows. In the next section, the CMU PDA speech database is introduced. It is also described how this database motivated us to develop a system that can handle "inhomogeneous" microphones. After a brief description of DCN in Section 3, Decoder-Based Channel Selection (DBCS) is proposed in Section 4. Experimental results are shown in Section 5, and the last section gives conclusions and future works.

2. CMU PDA speech database

Prior to this work, we had created the CMU (Carnegie Mellon University) PDA speech database to investigate applicability of various algorithms to Personal Digital Assistant (PDA) speech recognition [6]. In [1], we used the single-microphone version to evaluate DCN. In this work, we

Table 1: *Baseline word error rates.*

	PDA-A	PDA-B	Ave.	Rel. imp.
CH. 1	25.7	67.7	46.7	-
CH. 2	23.5	64.9	44.2	-
CH. 3	18.9	61.2	40.1	-
CH. 4	29.0	67.5	48.2	-
Ave.	24.3	65.3	44.8	0.0
D&S	22.6	62.7	42.6	4.8
Best	15.8	54.1	34.9	22.1
CLS	11.1	25.0	18.1	59.6

use the multiple-microphone version. This version of the database consists of two sets. The first set (PDA-A) was recorded in a rather quiet condition whose average SNR is estimated as 26dB. Each of eight speakers read 40-43 sentences (total 330 sentences) chosen from the LDC Wall Street Journal database (WSJ0). The speaker held a PDA (HP iPaq Pocket PC) so that the screen could be seen easily. The sentences appeared on the screen, and the utterances were recorded by the four microphones placed at the corners of a mockup that is attached to the PDA. These four microphones form a rectangle around the PDA, 5.5 cm across and 14.6 cm top-to-bottom. The second set (PDA-B), consisting of the same sentences uttered by eight other speakers in the same room, was recorded in a noisier condition with the average SNR estimated as 17dB. All the utterances were also recorded by the close-talking microphone worn by the speaker.

Table 1 shows the baseline word error rates (WERs) for the CMU PDA speech database. Throughout all experiments described in this paper, the Sphinx-III decoder developed by CMU [7] was used for decoding, with a trigram language model. The acoustic models were trained using the WSJ0 clean speech database. CMN is included in the baseline processing. The delay-and-sum beamformer consists of four steps: upsampling from 16kHz to 64kHz, beam steering by calculating the correlation to the first microphone within ±1ms, averaging with gain normalization, and downsampling from 64kHz to 16kHz. Single channel WERs (CH.1 to CH. 4) show that there is a large inhomogeneity among channels. The delay-and-sum beamformer (D&S) gives better WERs than the average of four single-channel WERs, but it is much worse than the best channel (CH. 3). Thus, if we know which channel is the best, we could achieve 10.5% relative improvement of the WER from the average, and 7.2% from the delay-and-sum beamformer. Moreover, if we have the “oracle” knowledge to choose the best channel for each utterance (shown as “Best”), we could obtain 22.1% relative improvement from the average. Although it is not obvious which channel should be chosen for each utterance, these results suggest that a good indicator of the channel quality is necessary to deal with inhomogeneous microphones.

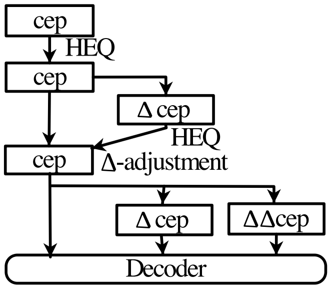

Figure 1: *Schematic diagram of DCN.*

Table 2: *A comparison of WERs obtained by single-channel feature compensation algorithms.*

	PDA-A	PDA-B	Ave.	Rel. imp.
Baseline (CMN)	24.3	65.3	44.8	0.0
MVN	20.4	59.5	40.0	10.7
HEQ	24.6	71.2	47.9	-6.9
DCN	21.9	57.3	39.6	11.6
D&S	22.6	62.7	42.6	4.8
D&S + MVN	18.6	56.7	37.6	16.0
D&S + HEQ	23.9	70.3	46.6	-4.1
D&S + DCN	20.9	56.0	38.4	14.2

3. Delta-Cepstrum Normalization

Delta-Cepstrum Normalization (DCN) [1] is a feature compensation algorithms that is effective in highly noisy conditions and could be implemented on small devices. Figure 1 shows the schematic diagram of DCN. First, the cepstrum is normalized using Histogram Equalization (HEQ). The delta-cepstrum is calculated using the normalized cepstrum. HEQ is then applied to the delta-cepstrum to obtain the better distribution of the delta-cepstral coefficients. After that, inconsistency between the normalized cepstrum and the normalized delta-cepstrum is reduced by the procedure called Δ-adjustment. Finally, the delta-cepstrum and the delta-delta-cepstrum are re-calculated using the output of the Δ-adjustment procedure, and all the feature vectors are fed into the decoder.

Table 2 shows the WERs obtained by various single-channel feature compensation algorithms. Each WER on the upper four lines represents the average of four channels. It is disappointing and opposite to the previous work that HEQ gives worse results than the baseline, but at any rate, DCN gives the best performance. It can be interpreted that the overfitting effect of HEQ was compensated by DCN. In particular, the improvement is greater for the noisy set (PDA-B) as expected. The lower four lines show the results obtained by compensating the output of the delay-and-sum beamformer using the same single-channel algorithms. It can be seen that combining the delay-and-sum beamformer and those single-channel compensation algorithms gives some additional improvements.

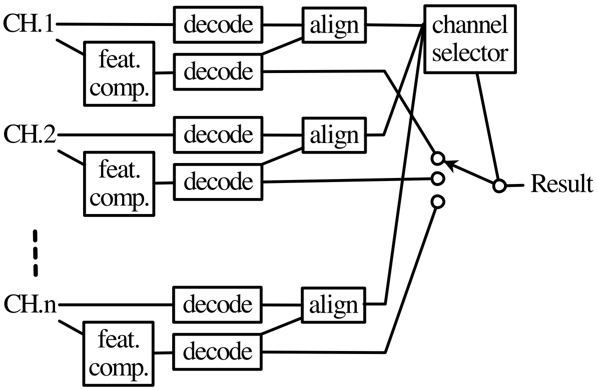

Figure 2: *Schematic diagram of DBCS.*

4. Decoder-Based Channel Selection

As the baseline experiments have shown, it is important to have a good indicator of the channel quality when the input signal consists of inhomogeneous multiple channels. If the inhomogeneity is caused by the input signal power, representing the difference of the length between the speaker's mouth and the microphone, the estimated SNR could be a good indicator. A more reliable indicator would be the likelihood given by the recognizer[8].

However, if there are other dominant factors, SNR-based indicators can not work effectively. Likelihood-based indicators are more reliable, but some confusing words may give high likelihood. Instead, we use the effectiveness of the single-channel feature compensation algorithm as the indicator of the channel quality. If the feature compensation gives more benefit, we assume that the original input is more contaminated.

The simplest way to evaluate the effectiveness of the feature compensation is to compare the hypotheses made with/without compensation. Two hypotheses are compared using a DP-based alignment program. Thus, the channel that had the fewest mismatched words is selected to be used for recognition. This algorithm, referred to as Decoder-Based Channel Selection (DBCS), could be implemented easily, and gives great improvement to the recognition accuracy. The schematic diagram of DBCS is shown in Fig. 2.

5. Experimental Results

The proposed algorithm was evaluated using the CMU PDA speech database. Table 3 shows the WERs obtained by various channel selection methods. In this experiment, MVN, HEQ, and DCN were used only for channel selection to evaluate the performance of channel selection. First, the NIST SPQA tool [9] was used to estimate the SNR of each utterance. The channel with the largest SNR was used for recognition, but the result was at the same level with the baseline. Second, the likelihood provided by the Sphinx-III decoder was used, and some WER improvement was obtained in this case. Finally, DBCS was applied with three single-channel feature compensation algorithms. One of the single-channel compensation algorithms (MVN, HEQ, or DCN) was applied to obtain the compensated feature vectors. Both compensated and uncompensated feature vectors were used to make the hypotheses for each channel, and the two hypotheses made from one channel were aligned. The channel with the fewest mismatched words was used for recognition. In this experiment, the uncompensated feature vectors were used to evaluate the performance of channel selection only, instead of using the compensated feature vectors to obtain better results. As shown in the table, 10-13% relative improvements were achieved by DBCS, which are much higher than likelihood-based channel selection. It should be noted that three kinds of DBCS provide similar results even though HEQ was not helpful as the single-channel feature compensation for this database.

Table 3: *A comparison of WERs obtained by various channel selection methods. MVN, HEQ, and DCN were used only for channel selection.*

	PDA-A	PDA-B	Ave.	Rel. imp.
SNR	25.3	64.3	44.8	0.0
Likelihood	20.7	63.6	42.1	6.0
MVN-DBCS	19.0	59.9	39.5	11.8
HEQ-DBCS	18.9	61.4	40.2	10.3
DCN-DBCS	19.2	58.5	38.9	13.2

Table 4: *A comparison of WERs obtained by DBCS in combination with single-channel feature compensation. The bottom line shows the ideal case where the best of 4 DCN outputs is known for each utterance.*

	PDA-A	PDA-B	Ave.	Rel. imp.
MVN-DBCS	16.3	55.6	35.9	19.9
HEQ-DBCS	18.7	64.2	41.4	7.6
DCN-DBCS	17.8	51.6	34.7	22.5
DCN-Likelihood	19.5	50.4	34.9	22.1
DCN-Best	14.2	42.9	28.5	36.4

Table 4 shows the results of the same experiment except that the compensated feature vectors were used after the channel selection. Obviously, using the feature vectors compensated by HEQ increases the WERs, that is consistent with the baseline experiment. In contrast, results of MVN-DBCS and DCN-DBCS are improved by using compensated feature vectors. The best case, DCN-DBCS, gives 23% relative improvement from the baseline, that is 19% relative improvement from the delay-and-sum beamformer. In these experiments, likelihood-based channel selection also gives good improvement, but DBCS is slightly better.

The concept of DBCS is also applicable to the output of the delay-and-sum beamformer. If some channels are reliable and some are not, the delayed sum of only the reliable channels is expected to be cleaner than any single channel or the

Table 5: *A comparison of WERs obtained by DBCS in combination with single-channel feature compensation. The best channel was selected from 4 single-microphone channels and 11 partial delay-and-sum beamformers.*

	PDA-A	PDA-B	Ave.	Rel. imp.
MVN-DBCS	15.8	52.7	34.2	23.5
HEQ-DBCS	16.6	60.3	38.4	14.2
DCN-DBCS	15.6	47.8	31.7	29.2
DCN-Likelihood	19.1	47.8	33.5	25.2
DCN-Best	11.2	35.1	23.2	48.3

delayed sum of all channels. Since we have four channels, there are eleven combinations of two, three, or four channels. Table 5 shows the WERs obtained using the single-channel inputs and those 'partial' beamformers. The WERs are reduced in all cases. In particular, DCN-DBCS gives excellent performance, which is 29% relative improvement from the baseline, 26% from the delay-and-sum beamformer, and 5.4% from likelihood-based channel selection.

The problem of the proposed algorithm is the computational cost. It requires two decoding runs for each channel, resulting in eight runs in the four channel case, or thirty runs if we apply it to the partial beamformers. One solution for this problem is to use the best channel that is selected for the preceding utterance. We carried out another experiment, in which the best channel is selected using one utterance, and the result is used for N utterances. The results are shown in Fig. 3. Obviously, a large gap exists between $N = 1$ and $N = 2$, which means that there is an environmental factor that varies utterance by utterance. For $N > 1$, WER approaches asymptotically to the value of the delay-and-sum beamformer. However, some improvements can still be obtained at $N = 10$ or $N = 15$.

6. Conclusions

In this paper, the importance of channel quality estimation was introduced. If multiple microphones are not homogeneous, due to thier relative position or other reasons, the output of the standard delay-and-sum beamformer is better than the average of all channels, but worse than the best channel. Therefore, it is important to find a good indicator to select a channel for recognition. We proposed Decoder-Based Channel Selection (DBCS), in which the hypotheses made by compensated and uncompensated feature vectors were aligned to evaluate the effectiveness of the single-channel feature compensation. Using the channel that has the fewest mismatched words, the WER can be greatly reduced. DBCS with Delta-Cepstrum Normalization (DCN) for single-channel compensation gave 26% relative WER reduction from the standard delay-and-sum beamformer if we use all single channels plus all possible combinations of them.

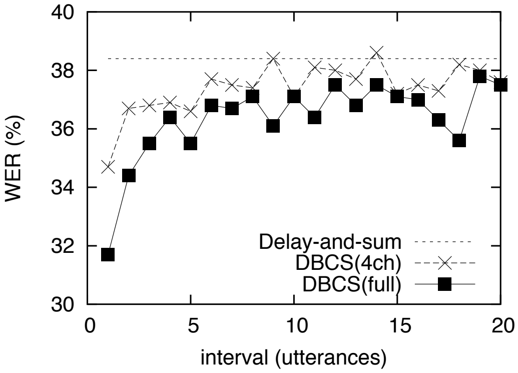

Figure 3: *WERs obtained with various intervals of channel selection. DCN was applied in all cases.*

DBCS is computationally expensive because it requires two decoding runs for each channel during channel selection. Processing time can be reduced if we use the selection result for the preceding utterance, although the improvement becomes smaller. Another approach, which is one of our future works, would be to use a simpler decoder for channel selection while the full decoder is used for the second path. Omitting the language model scoring could reduce the processing time, as much as replacing the HMM based decoder with DTW or GMM based decoders.

7. References

[1] Y. Obuchi and R. M. Stern, "Normalization of time-derivative parameters using histogram equalization," *Proc. EUROSPEECH*, Geneva, Switzerland, 2003

[2] A. de la Torre, et al., "Non-linear transformation of the feature space for robust speech recognition," *Proc. ICASSP*, Orlando, USA, 2002

[3] B. S. Atal, "Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification," *Journal of the Acoustical Society of America*, vol.55, pp.1304-1312, 1974

[4] W. Kellermann, "A self steering digital microphone array," *Proc. ICASSP*, Toronto, Canada, 1991

[5] J. P. Openshaw and J. S. Mason, "On the limitations of cepstral features in noise," *Proc. ICASSP*, Adelaide, Australia, 1994

[6] M. L. Seltzer, "*Microphone array processing for robust speech recognition*," Ph.D thesis, Carnegie Mellon University, Pittsburgh, USA, 2003

[7] http://www.speech.cs.cmu.edu/speech/sphinx/

[8] Y. Shimizu, et al., "Speech recognition based on space diversity using distributed multi-microphone," *Proc. ICASSP*, Istanbul, Turkey, 2000

[9] http://www.nist.gov/speech/tools/