Composite Decision by Bayesian Inference in Distant-Talking Speech Recognition

Mikyong Ji, Sungtak Kim, and Hoirin Kim

SRT Lab., Information and Communications University
119, Munjiro, Yuseong-gu, Daejeon, 305-732, Korea
{lindaji, stkim, hrkim}@icu.ac.kr

Abstract. This paper describes an integrated system to produce a composite recognition output on distant-talking speech when the recognition results from multiple microphone inputs are available. In many cases, the composite recognition result has lower error rate than any other individual output. In this work, the composite recognition result is obtained by applying Bayesian inference. The log likelihood score is assumed to follow a Gaussian distribution, at least approximately. First, the distribution of the likelihood score is estimated in the development set. Then, the confidence interval for the likelihood score is used to remove unreliable microphone channels. Finally, the area under the distribution between the likelihood score of a hypothesis and that of the $(N+1)^{st}$ hypothesis is obtained for every channel and integrated for all channels by Bayesian inference. The proposed system shows considerable performance improvement compared with the result using an ordinary method by the summation of likelihoods as well as any of the recognition results of the channels.

1 Introduction

The state-of-the-art speech recognizers can achieve high recognition accuracy when close-talking microphones are used. However, in distant-talking environments, the performance is significantly degraded due to a variety of causes such as the distance between the sound source and the microphone, the position of the microphone, the direction of the speaker, the quality of the microphone, etc.

To cope with these problems, microphone array-based speech recognizers have been widely applied to improve not only the quality of the speech but also the recognition performance [1,2,3]. The simplest beamforming processing using the delay-and-sum principle has been successfully used. However, it is difficult to estimate time-delay accurately in noisy and reverberant environments [4]. While the use of the microphone array can capture only one-directional acoustic information, the use of the spatially distributed multiple microphones can capture spatial acoustic information in a room [5]. In addition, the latter makes the arrangement of microphones in a room easier than the former.

In this paper, we propose methods to improve the performance of the distant-talking speech recognition by producing a composite decision from the recognition results with multiple microphones. In the work, the distribution of the likelihood score is assumed to be a Gaussian density function. Its distribution is estimated, the confidence interval for the likelihood score is found to extract unreliable results, and the area under the distribution

Petr Sojka, Ivan Kopeček and Karel Pala (Eds.): TSD 2006, LNAI 4188, pp. 463–470, 2006.

between the likelihood score of the hypothesis and that of the $(N + 1)^{st}$ hypothesis is computed per channel. Eventually, it is merged into a composite result by Bayesian inference.

The remainder of the paper is organized as follows. In Section 2, we introduce methods to combine the recognition results by simultaneously recorded speech inputs into a composite one, and Section 3 describes the experimental results and the performance evaluation. Finally, we conclude and describe future works in Section 5.

2 Integration by Bayesian Inference

Speech inputs recorded simultaneously through two or more microphones are separately recognized per recognizer and their results are combined into the best scoring hypothesis by the integration module in Fig. 1. Given the speech inputs, $X_1, X_2, \ldots, X_C$ obtained by multiple microphones, the best hypothesis should be chosen to maximize $P(W|X_1, X_2, \ldots, X_C)$. If we assume that the speech inputs from different channels are conditionally independent given a hypothesis and each hypothesis has an equal prior probability $P(W) = 1$, this can be further simplified by Bayesian inference as

$$\overline{W} = \arg\max_W p(W|X_1, X_2, \cdots, X_C) = \arg\max_W \prod_{i=1}^{C} p(X_i|W) \tag{1}$$

where C is the number of microphones.

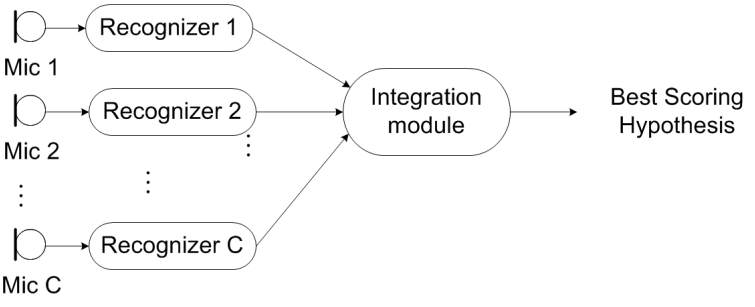

Fig. 1. System architecture

The relation between the best hypothesis and the multiple speech inputs is described into a Bayesian network shown in Fig. 2.

2.1 Integration by Likelihoods

If we take the logarithm of Eq. (1), which is simplified by Bayesian inference, the equation is represented as follows:

$$\overline{W} = \arg\max_W \sum_{i=1}^{C} \log p(X_i|W). \tag{2}$$

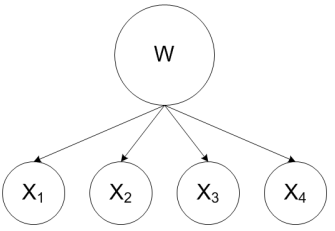

Fig. 2. Bayesian Network

The best hypothesis can be determined by the sum of the log likelihoods whose hypotheses are the same among the recognition results by multiple channels. That is, the best rescoring hypothesis is identified as a composite recognition result.

2.2 Reliable Channel Selection by Confidence Interval

The result of measurements is often accompanied by a confidence interval to determine an interval that has a high probability of containing the true value [6,7]. Thus, if we say we are (1-α)•100% confident between $-l_0$ and l_0 for the parameter l, it is described by

$$P(-l_o < l < l_o) = 1 - 2\alpha \tag{3}$$

where α is a number between 0 and 1. In other words, there is only an α% chance that l will be less than $-l_0$ and an α% chance that will be larger than l_0. In this paper, the confidence interval for the log likelihood score of the hypothesis, $\log p(W|X_c)$ is computed. The lower limit is employed to remove an unreliable channel of which the likelihood of the best hypothesis lies under the upper tail area in Fig. 3.

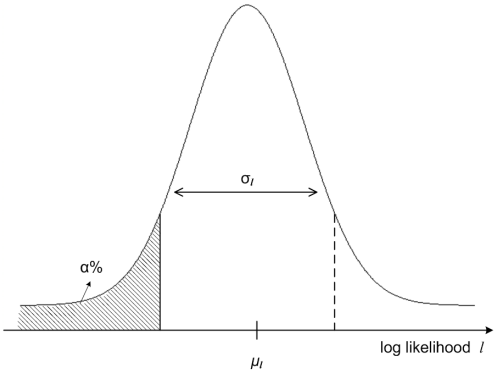

Fig. 3. Confidence interval

The log likelihood score $log\ p(W|X_c)$ is assumed to follow a Gaussian distribution with mean μ_l and variance σ_l^2. Afterwards, the distribution is estimated. The detail about the

estimation will be described in the following section. Since there is no closed form to integrate the Gaussian PDF, the bound is found by transforming into the standard Gaussian function and using its table in Eq. (4).

$$P\left(\frac{-l_o-\mu_l}{\sigma_l} < \frac{l-\mu_l}{\sigma_l} < \frac{l_o-\mu_l}{\sigma_l}\right) = 1-2\alpha. \tag{4}$$

2.3 Integration Using Area Under Density Curve Between Hypotheses' Scores

We assume that the log likelihood score l follows a Gaussian distribution and the conditional observation distribution of $l|\mu$ is Gaussian with mean μ and variance σ^2, which is assumed known. Then, its density is as follows:

$$p(l|\mu) \propto e^{-\frac{1}{2\sigma^2}(l-\mu)^2}. \tag{5}$$

The part that doesn't depend on the parameter μ is the same for all parameter values; if we ignore the constant of proportionality, it can be represented by Eq. (5). Suppose that the prior probability for the parameter μ is a flat prior density ($f(\mu) = 1$). The shape of the posterior for μ is given by

$$p(\mu|l) \propto p(\mu) \cdot p(l|\mu) \propto e^{-\frac{1}{2\sigma^2}(\mu-l)^2}. \tag{6}$$

On the other hand, if we have the Gaussian distribution with mean m and variance s^2 for μ, the shape of the posterior is

$$p(\mu|l) \propto e^{-\frac{1}{2\sigma^2 s^2/(\sigma^2+s^2)}\left(\mu-\frac{(\sigma^2 m+s^2 l)}{\sigma^2+s^2}\right)^2}. \tag{7}$$

The update of the PDF of μ can be simplified by Eq. (8).

$$m' = \frac{\sigma^2}{\sigma^2+s^2}\times m + \frac{s^2}{\sigma^2+s^2}\times l, \quad (s')^2 = \frac{\sigma^2 s^2}{\sigma^2+s^2}. \tag{8}$$

Consequently, the distribution for the next observation l_{n+1} is described by

$$\begin{aligned} p(l_{n+1}|l_1, l_2, \cdots, l_n) &= \int p(l_{n+1}, \mu|l_1, l_2, \cdots, l_n)d\mu \\ &= \int p(l_{n+1}|\mu) \times p(\mu|l_1, l_2, \cdots, l_n)d\mu \\ &\propto e^{-\frac{1}{2(\sigma^2+s_n^2)}(l_{n+1}-m_n)^2} \end{aligned} \tag{9}$$

where we are ignoring the part that does not involve μ.

$$p(l) \propto e^{-\frac{1}{2(\sigma^2+s^2)}(l-m)^2}. \tag{10}$$

Instead of using the log likelihood score of the hypothesis itself from the recognition result directly as in Section 2.1, the proposed method estimates the distribution of the likelihood

score, $p(l)$ and the area under the distribution between the log likelihood score of the hypothesis and the $(N+1)^{st}$ hypothesis is computed in Eq. (11).

$$A^c(W) = P(L^c(W^{N+1}) < l < L^c(W)) = \int_{L(W^{N+1})}^{L(W)} p(l)dl \tag{11}$$

where l is the log likelihood score, W^{N+1} is the $(N+1)^{st}$ hypothesis in the N-best list, $L^c(W^{N+1})$ is the log likelihood score of the $(N+1)^{st}$ hypothesis given the speech input by microphone c, X_c, and $p(l)$ is the PDF of the likelihood score. That is, the area under the distribution between the log likelihood score of a hypothesis and that of the $(N+1)^{st}$ hypothesis is computed per channel input and it is integrated into Bayesian inference introduced in Eq. (1). The composite result is decided by Eq. (12).

$$\overline{W} = \arg\max_{W} \sum_{c=1}^{C} A^c(W). \tag{12}$$

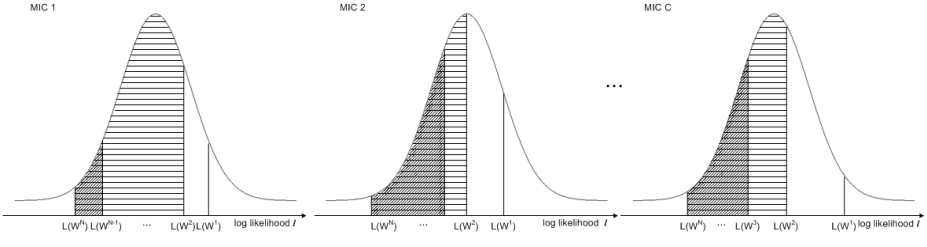

Fig. 4. Area under the distribution between the log likelihood of each hypothesis and that of the $(N+1)^{st}$ hypothesis for each microphone input

3 Experiments

3.1 Experimental Setup

For the experiments, we use Korean POW (phonetically optimized word) 3848 database. It consists of 3848 different words which are divided to 8 sub-sets and each speaker uttered one of 8 sub-sets. The total number of speakers is 80 (40 males and 40 females). To show the effectiveness of the proposed system, only 268 words among the database are selected. As shown in Fig. 5, the selected 268 words are recorded again by using a loudspeaker at 5 positions marked in a room where four microphones are installed at the four corners to face its center. The speech inputs through four microphones were sampled by 16 KHz. Also, the five kinds of music without vocal sound were recorded by four channels in the same environment in Fig. 5 and added per channel to make noisy database.

We use MFCCs, their corresponding delta and acceleration coefficients as the feature vectors. A pre-emphasis filter $H(z) = 1 - 0.97z^{-1}$ is used before framing and each frame is multiplied with a 20 ms Hamming window, shifted by 10 ms.

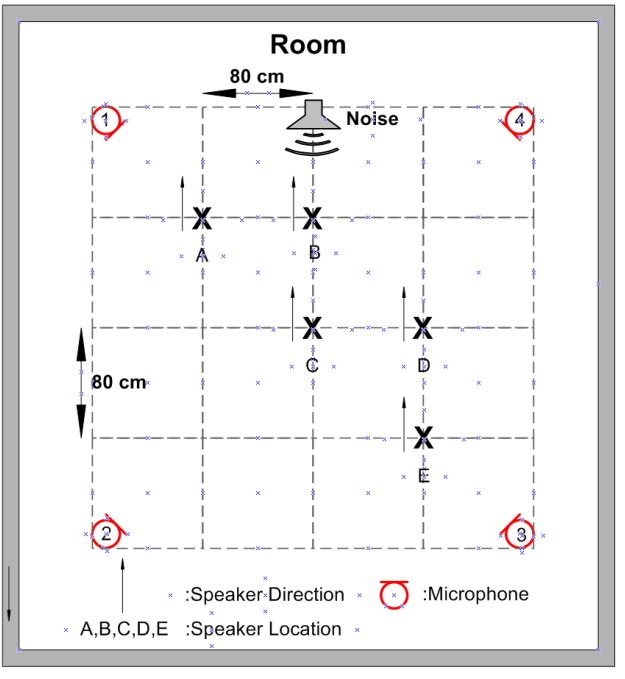

Fig. 5. Environment of DB Collection

3.2 Experimental Results

For performance comparison, we applied four different integration methods to multiple recognition results: integration by the magnitude of the likelihood (ML), by the summation of the likelihoods (BN_L), by the area under the likelihood density curve (BN_D), and by applying a confidence interval in order to remove unreliable channels ahead of the BN_D (BN_DC). Table 1 shows the baseline recognition results by the speech inputs through four multiple microphones. Table 2 and 3 represent the recognition results per location and they are followed by four different composite recognition results. Table 4 describes the error reduction rate (ERR) of BI_DC over other integration methods. ORACLE is the maximum reachable word accuracy. It describes whether the correct word is included in the N-best lists for integration.

Table 1. Recognition accuracy of the baseline system (%)

CH SNR	MIC1	MIC2	MIC3	MIC4
Clean	88.35	73.57	83.86	89.67
10 dB	81.49	83.33	82.28	82.33
5 dB	60.08	66.98	66.19	61.89
0 dB	22.72	32.40	30.63	23.98

As shown in the experimental results, the composite recognition result is improved compared with an individual output by each microphone. The performance improvement was significant when the integration by the area under the likelihood density curve was applied and it was applied after removing unreliable channels with a confidence interval even more. As N got increased, the Word Error Rate was decreased by only small amount.

Table 2. Performance comparison of integration methods (Clean, N=1, %)

LOC CH	LOCA	LOCB	LOCC	LOCD	LOCE	AVG
MIC1	89.29	88.81	91.59	85.85	86.23	88.35
MIC2	73.23	66.63	86.52	70.27	71.22	73.57
MIC3	77.34	84.03	85.18	86.42	86.33	83.86
MIC4	86.71	90.44	91.11	91.59	88.53	89.67
ML	85.47	87.28	88.24	90.25	88.81	88.01
BI_L	87.28	89.01	91.87	89.58	88.43	89.23
BI_D	90.25	90.73	91.40	91.30	91.01	90.94
BI_DC	90.44	90.73	91.40	91.59	91.11	91.05
ORACLE	95.89	95.89	96.85	96.37	95.79	96.16

Table 3. Performance comparison of integration methods (Music noise 5 dB, N=1, %)

LOC CH	LOCA	LOCB	LOCC	LOCD	LOCE	AVG
MIC1	64.05	56.50	62.43	59.18	58.22	60.08
MIC2	72.08	64.44	69.02	66.06	63.29	66.98
MIC3	69.50	63.67	68.45	65.87	63.48	66.19
MIC4	66.44	59.75	62.43	61.66	59.18	61.89
ML	72.47	63.96	68.64	65.68	64.63	67.08
BI_L	71.89	65.49	68.36	66.54	64.15	67.29
BI_D	74.19	67.50	71.32	67.78	65.68	69.29
BI_DC	74.67	67.69	71.70	67.88	65.58	69.50
ORACLE	82.03	77.06	80.50	77.25	76.00	78.57

Table 4. ERR of BI_DC over other integration methods (%)

Method SNR	Best Channel	ML	BI_L	BI_D
Clean	21.26	37.30	26.26	2.11
10 dB	16.07	6.85	12.13	1.32
5 dB	21.74	21.06	19.59	2.26
0 dB	-1.30	22.01	16.38	1.27

4 Conclusion

The integrated system to produce a composite speech recognition output has been proposed and it is shown that the integration of the recognition results from spatially distributed

microphones is effective in distant-talking speech recognition. After we assume that the likelihood score follows a Gaussian distribution, the area under the distribution between hypotheses is computed and combined into the best scoring hypothesis. When this proposed method is applied after removing unreliable channels, the best performance is achieved. However, the distribution of the likelihood score should be estimated in advance; it is still useful from the viewpoint that it can greatly contribute to the performance improvement in distant-talking speech recognition to realize hands-free applications.

In this paper, we considered one-directional noise source, and restricted the number of distributed microphones to four. Thus, diverse experiments are required to confirm the effectiveness of our system.

References

1. Hughes, T. B., Kim, H., Dibiase, J. H., Silverman, H. F.: Using A Real-Time, Tracking Microphone Array as Input to an HMM Speech Recognize. Proc. of ICASSP, Vol. 1. (1998) 249–252.
2. Yamada, T., Makamura, S., Shikano, K.: Hands-free Speech Recognition with Talker Localization by a Microphone Array. Trans. of Information Processing Society of Japan, Vol. 39. No. 5. (1998) 1275–1284.
3. Takiguchi, T., Nakamura, S., Shikano, K.: HMM-Seperation Based Speech Recognition for a Distant Moving Speaker. IEEE Trans. on Speech and Audio Processing, Vol. 9. No. 2. (2001) 127–140.
4. Yamada, T., Nakamura, S., Shikano, K.: Distant-Talking Speech Recognition Based on a 3-D Viterbi Search Using a Microphone Array. Trans. on Speech and Audio Processing, Vol. 10. No. 2. (2002) 48–56.
5. Shimizu, Y., Kajita, S., Takeda, K., Itakura, F.: Speech Recognition Based on Space Diversity Using Distributed Multi-Microphone. Proc. of ICASSP, Vol. 3. (2000) 1747–1750.
6. Bolstad, W. M.: Introduction to Bayesian Statistics. John Wiley & Sons, Hoboken New Jersey (2004).
7. Neapolitan, R. E.: Learning Bayesian Networks. 2nd edn. Pearson Prentice Hall, Upper Saddle River New Jersey (2004).