P3-23

Text-Independent Speaker Identification using Soft Channel Selection in a Multi-Microphone Environment

Mikyong Ji, Sungtak Kim, Hoirin Kim , *Member*, *IEEE*, and Ho-Sub Yoon

***Abstract*—With the aim of improving speaker identification in a multi-microphone environment, we develop a text-independent speaker identification system. It incorporates soft channel selection before the combination of the identification results obtained by multiple microphones. The results demonstrate that the proposed system achieves high classification accuracy, thereby providing a speech interface for a wide range of potential hands-free applications in a ubiquitous environment.**

I. INTRODUCTION

State-of-the-art speaker identification (SI) technologies have achieved high recognition accuracy. Even if one of the current technologies would yield the best identification rate, its performance could be significantly degraded due to a variety of causes in a distant-talking environment. To deal with it, microphone array-based speaker recognizers have been successfully applied through speech enhancement [1], [2]. But, the accurate estimation of the time delays between the different speech signals is still not an easy task due to background noise, room reverberation, the non-stationary characteristics of the speech signal, etc. There has been also another approach based on feature compensation for robust SI in a multi-microphone environment [3]. In this paper, we propose a new SI system, which can greatly enhance the identification rate by combining the identification results with soft channel selection with a single perceptron in a multi-microphone environment.

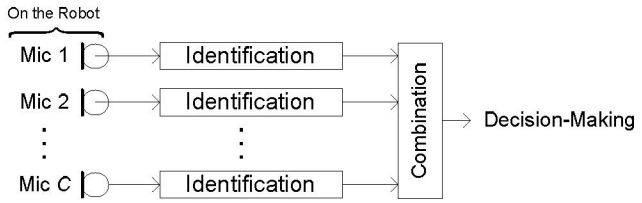

Fig. 1. Combining the identification results with multiple microphones in a distant-talking environment.

II. SPEAKER IDENTIFICATION WITH MULTIPLE MICROPHONES

A. Combining Speaker Identification Results

Given different speech inputs $X_1, X_2, \ldots, X_C$ simultaneously recorded through C multiple microphones, the speaker who provides $X_1, X_2 \ldots, X_C$ among a set of known speakers $\mathcal{S} = \{1, 2, \ldots, S\}$ is generally identified by (1). Each speaker is modeled individually by Gaussian mixture model (GMM) λ_k.

$$\hat{S} = \arg\max_{1\le k\le S} P(\lambda_k \mid X_1, X_2, \ldots, X_C). \tag{1}$$

This work is a result of the URC project sponsored by the MIC of the Korean government.

Depending on the assumption made, (1) can be rewritten as one of the following combination rules, CS (combination by sum), CM (combination by max), and CV (combination by voting) as in (2), (3) and (4), respectively [4].

$$\hat{S} \cong \arg\max_{1\le k\le S} \sum_{c=1}^{C} \log P(\lambda_k \mid X_c). \tag{2}$$

$$\hat{S} \cong \arg\max_{1\le k\le S} \max_{1\le c\le C} P(\lambda_k \mid X_c). \tag{3}$$

$$\hat{S} \cong \arg\max_{1\le k\le S} \sum_{c=1}^{C} \Delta_{kc}, \tag{4}$$

where Δ_{kc} is further defined by

$$\Delta_{kc} = \begin{cases} 1 & if\ P(\lambda_k \mid X_c) = \max\limits_{1\le k\le S} P(\lambda_k \mid X_c) \\ 0 & otherwise \end{cases}. \tag{5}$$

B. Entropy by Posterior Probabilities

The entropy $H(Y)$ of discrete random variable $Y = \{y_1, y_2, \ldots, y_N\}$ introduced by Shannon is defined as:

$$H(Y) = -\sum_{i=0}^{N} P(y_i) \log_b P(y_i), \tag{6}$$

where $P(y_i) \ge 0$, $\sum_i P(y_i) = 1$, and $P(y_i) = Pr(Y=y_i)$. The entropy should be maximal if all the outcomes are equally likely ($P(y_i) = 1/N$). For all N, it follows

$$H(P(y_1), P(y_2), \ldots, P(y_N)) \le H(\frac{1}{N}, \frac{1}{N}, \ldots, \frac{1}{N}) = \log_b^N . \tag{7}$$

Let us assume that there exist S enrolled speakers, and they are equally likely. Then, the posterior probability of speaker k at frame t is represented as follows:

$$p_k^t = P(\lambda_k \mid x_t) = \frac{p(x_t \mid \lambda_k)}{\sum_{k=1}^{S} p(x_t \mid \lambda_k)}. \tag{8}$$

In this paper, the posterior probabilities of the individual speakers, $p_1^t, p_2^t, \ldots, p_S^t$, are employed to compute the entropy at frame t. The entropy should be maximal in the same manner as (7) if the posterior probabilities are equal ($p_k^t = 1/S$) for all k.

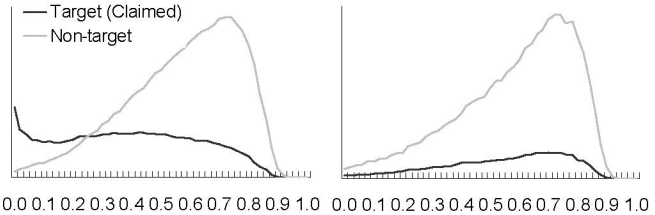

Fig. 2. Distribution of the frame's entropy in the case of speaker identification success (left) and failure (right).

C. Identified Speaker's Continuity

Frame-pruning technologies have been previously applied

to improve the identification rate by removing certain frames that do not really contribute to identifying the actual speaker using a divergence measure [5]. The entropy is used to measure the degree of contribution to identifying the speaker. However, they resulted in relatively low performance since they remove the only frames whose entropies are large, that is, which do not greatly affect the classification result. Figure 2 shows the frame's entropy which the target (claimed) or the non-target speaker is recognized in the case of identification success and failure, respectively. The frames which the target speaker is recognized on are relatively more distributed over low value in the case of correct identification than incorrect identification. Thus, we propose a feature called an identified speaker's continuity (ISC), which represents the confidence in the identified speaker. Pseudo code for ISC is described in Fig. 4.

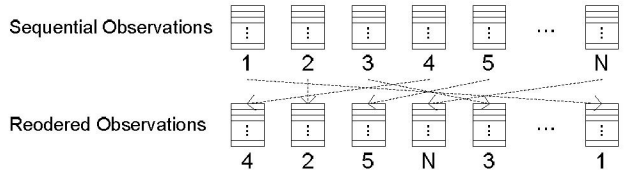

Fig. 3. Observation sequence reordered by entropy in ascending order.

```
Identify a speaker (target or claimed speaker s')
Initialize a counter to 0
Reorder observation sequence by the entropy in ascending order as in Fig. 3
For each reordered frame n (Loop1)
    For each speaker s (Loop2)
        Evaluate Σ_{i=0}^{n} p(x_i|λ_s)
    End (Loop2)
    Find the speaker m with the maximimum
    if (m=s') increment the counter
    else initialize the counter to 0
End (Loop1)
Divide the counter by the total number of frames
```

Fig. 4. Pseudo code for identified speaker's continuity (ISC).

D. Soft Channel Selection by using Perceptron

By selecting the only reliable channels and combining the results obtained from them, the speaker identification can be improved even further. As shown in Fig. 5, a single perceptron learned by gradient descent algorithm is used for soft channel selection. The ISC and voting [6] are used as the inputs to the perceptron.

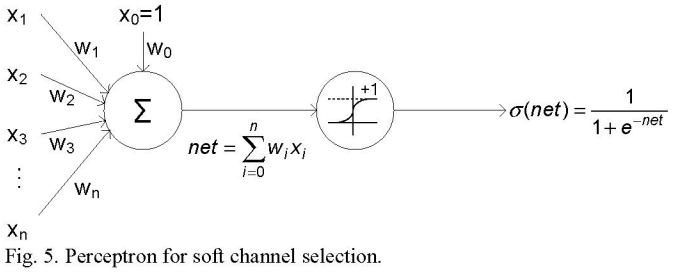

Fig. 5. Perceptron for soft channel selection.

III. Experiment

The evaluation was performed with a database uttered by 30 speakers (23 males and 7 females). Nearly 60 conversational sentences, with lengths of about one to two seconds, were recorded in a quiet environment by each speaker. They were then re-recorded again with eight microphones on the robot by playing them back with a loudspeaker placed at center (0°) or diagonal (45°) with distances of 1m, 3m, and 5m and facing the robot (mock-up) in a home environment. Among them, nearly 30 different sentences per speaker, each of which was recorded at center or diagonal 1m by eight microphones, were used to train GMMs and the perceptron. The rest of them were used for evaluation. Figure 6 shows the identification accuracy of combination rules before and after the application of soft channel selection. The proposed system's performance gets better as the distance increases.

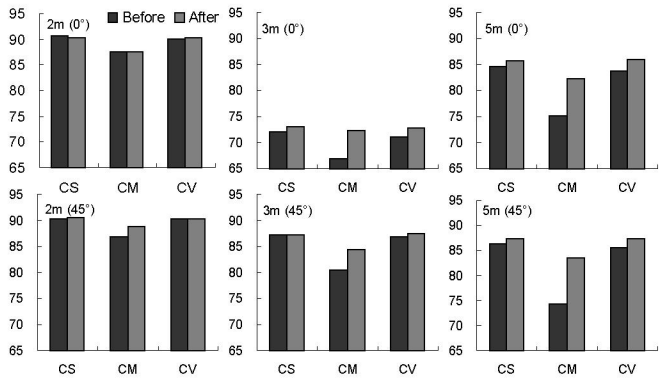

Fig. 6 Various combination rules before and after the application of soft channel selection.

IV. Conclusion

We proposed a SI system, which can greatly enhance the quality of human and computer interaction by integrating the identification results using soft channel selection with a single perceptron in a distant-talking multi-microphone environment. The results demonstrated that the proposed system improves the identification performance even more when the speaker is somewhat away from the microphone. That is, it is suggested that the proposed system can be employed to advance the performance of distant-talking SI in a wide range of potential hands-free applications.

Reference

[1] Q. Lin, E Jan, and J. Flanagan,, "Microphone arrays and speaker identification," *IEEE Trans. Speech, Audio Processing*, vol. 2, no. 4, pp. 622-629, Oct. 1994.

[2] I. A. McCowan, J. Pelecanos, and S. Sridharan, "Robust speaker recognition using microphone arrays," in *Proc. Speaker Odyssey: The Speaker Recognition Workshop* (*Odyssey 2001*), Crete, Greece, 2001, pp. 101-106.

[3] Q. Jin, Y. Pan, and T. Schultz, "Far-field speaker recognition," in *Proc. Int. Conf. on Acoustics, Speech, and Signal Processing* (*ICASSP2006*), Toulouse, France, 2006, pp. 937-940.

[4] J. Kittler, M. Hatef, R. P. W. Duin, and J. Matas, "On combining classifiers," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 20, no. 3, pp. 226-239. 1998.

[5] L. Besacier, and J. F. Bonastre, "Frame pruning for automatic speaker identification," in *Proc.* Int. *Conf. on European Signal Processing* (EUSIPCO'98), Island of Rhodes, Greece, 1998, pp. 367–370.

[6] B. Narayanaswamy and R. Gangadharaiah, "Extracting additional information from Gaussian mixture model probabilities for improved text-independent speaker identification," in *Proc. Int. Conf. on Acoustics, Speech, and Signal Processing* (*ICASSP2005*), Philadelphia, USA, 2005, vol. 1, pp. 621-624.