Signal-to-Signal Ratio Independent Speaker Identification for Co-Channel Speech Signals

R. Saiedi[1]*, P. Mowlaee[2], T. Kinnunen[1], Z. -H. Tan[2], M. G. Christensen[3], S. H. Jensen[2], and P. Fränti[1]

[1]*Dept. of Computer Science and Statistics, University of Eastern Finland, P.O. BOX 111, 80110 Joensuu, Finland*

[2]*Dept. of Electronic Systems, Aalborg University, Niels Jernes Vej 12, 9220 Aalborg, Denmark*

[3]*Dept. of Media Technology, Aalborg University, Niels Jernes Vej 14, 9220 Aalborg, Denmark*

{*rahim,tkinnu,franti*}*@cs.joensuu.fi*, {*pmb,zt,shj*}*@es.aau.dk, mgc@imi.aau.dk*

Abstract

In this paper, we consider speaker identification for the co-channel scenario in which speech mixture from speakers is recorded by one microphone only. The goal is to identify both of the speakers from their mixed signal. High recognition accuracies have already been reported when an accurately estimated signal-to-signal ratio (SSR) is available. In this paper, we approach the problem without estimating SSR. We show that a simple method based on fusion of adapted Gaussian mixture models and Kullback-Leibler divergence calculated between models, achieves an accuracy of 97% and 93% when the two target speakers enlisted as three and two most probable speakers, respectively.

1 Introduction

Speaker identification (SID) is the task of recognizing one's identity based on observed speech signal [1]. Typical speaker identification systems consist of short-term spectral feature extractor (front-end) and a pattern matching module (back-end). In traditional SID, the basic assumption is that only one target speaker exists in the given signal whereas in *co-channel* SID, the task is to identify two target speakers in one given mixture. Distinct from the so-called *summed channel* speaker recognition task [2], where only one speaker is talking most of the time, in the co-channel SID problem, both speakers talk simultaneously. Research on co-channel speaker identification has been done for more than one decade [3], yet the problem remains largely unsolved.

Most of the current *single-channel speech separation* (SCSS) systems use a model-based SID module, known as *Iroquois* [4] to identify the speakers in a mixed signal. The goal of an SCSS system is to estimate the unknown speaker signals according to their observed mixture. Interaction of the SID and speech separation modules can be managed in a closed loop to increase the overall performance [5]. Recognition accuracy as high as 98% has been reported for *Iroquois* in [6] which makes it as a first choice to be included in SCSS systems [7]. The database in [6] is provided for speech separation challenge and consists of 2 seconds of small vocabulary speech for 34 speakers. In the *Iroquois* system, a short list of the most likely speakers are produced based on the frames of the mixed signal that are dominated by one speaker. This short-list is then passed to a *max-based EM algorithm* to find the signal-to-signal ratio (SSR) and two speakers identity with an exhaustive search on codebooks created for speech synthesis [4].

The SSR estimation in *Iroquois* system is based on finding the most likely combination of speakers codebooks to produce the current speech frame, where in text-independent case gets more challenging compared to the database in [6]. Compared to the levels of $\{-9, -6, -3, 0, 3, 6\}$ dB SSRs considered by the *Iroquois* system, SSR can be continuous and time-varying over a recording in realistic conditions, making the SSR estimation even more difficult. Furthermore, in real-time applications of SCSS and in forensic applications it is necessary to have a *fast* and *accurate* system to identify the underlying sources in mixed signal without SSR estimation required.

To this end, in this paper, we propose an SSR-independent SID module for co-channel speech. More specifically, we examine different frame-level likelihood scores and model level distances to solve the problem and propose a combination of the most successful ones to compare the accuracy with repect to *Iroquois*. Since the proposed system is SSR-independent and tuned on 8 kHz speech, it is believed that it could be an alternative approach for the SID in SCSS and useful for telephony data found, for instance, in forensic applications.

*The work of R. Saeidi was supported by a scholarship from the Finnish Foundation for Technology Promotion (TES).

2 Speaker recognition approach

We use two main approaches for speaker recognition: frame-level log-likelihood calculation for a given mixed signal against a speaker GMM and between-models distance of a GMM model trained on mixed signal to speaker GMMs.

2.1 Frame-level likelihood scores

From the frame-level likelihood estimation originally defined for the *Iroquois* system in [4, 8] and which aims at determining the frames where only one speaker exists, we derive three different scores defined at the end of this section. A maximum likelihood (ML) trained GMM has been used in [4]; however, maximum *a posteriori* (MAP) derived GMMs [9] are more accurate in speaker verification and we follow this latter approach. Let λ denote speaker GMM. The likelihood function is defined as,

$$\ell(\mathbf{x}) = p(\mathbf{x}|\lambda) = \sum_{m=1}^{M} w_m p_m(\mathbf{x}). \tag{1}$$

The density is a weighted linear combination of M unimodal Gaussian densities $p_m(\mathbf{x})$, where $p_m(\mathbf{x}) \sim \mathcal{N}(\mathbf{x}; \mu_m, \mathbf{\Sigma}_m)$ and the mixture weights w_m further satisfy the constraints $\sum_{m=1}^{M} w_m = 1$ and $w_m \geq 0$. Speaker-dependent GMMs are adapted from universal background model (UBM) [9]. The UBM is a GMM trained on a pool of feature vectors extracted from as many speakers as possible to serve as *a priori* information for feature distribution. GMM means are the only parameters updated and weights and covariances are copied directly from UBM to GMMs.

2.2 Model distance scores

We define λ_{ig} as the SSR-dependent model for ith speaker at SSR level g. Another approach to measure similarity of a speech segment with a speaker model (λ_i) is to make a model from the test utterance with MAP adaptation (λ_e) and calculate the distance between λ_e and the speaker model. We use the *Kullback-Leibler divergence* (KLD) as a distance measure between the two probability distributions. Since this distance cannot be directly evaluated for GMMs, we use the upper bound of KLD which has successfully been applied to speaker verification [10]:

$$\begin{aligned}\mathrm{KLD}_i = \frac{1}{2}\sum_{g=1}^{G}\sum_{m=1}^{M} w_m(\boldsymbol{\mu}_{me} - \boldsymbol{\mu}_{mig})^T \mathbf{\Sigma}_m^{-1} \\ (\boldsymbol{\mu}_{me} - \boldsymbol{\mu}_{mig}).\end{aligned} \tag{2}$$

Here G ranges in a set of SSR levels, $\boldsymbol{\mu}_{me}$ is the mth mean vector in λ_e and $\boldsymbol{\mu}_{mig}$ is the mth mean vector in λ_{ig}, whereas w_m and $\mathbf{\Sigma}_m$ are the weights and the covariances of the UBM, respectively. An alternative approach to measure the distortion between GMMs is *approximate cross entropy* (ACE) [11]. As shown in [11], assuming infinite number of test utterance feature vectors, log-likelihood for a given λ_i equals to negative cross entropy between λ_e and λ_i. It can be approximated as follows:

$$\begin{aligned}\mathrm{ACE}_i = \sum_{g=1}^{G}\sum_{m=1}^{M} w_m \max_n \Big[\log w_n \\ - \frac{1}{2}(\boldsymbol{\mu}_{me} - \boldsymbol{\mu}_{nig})^T {\mathbf{\Sigma}_m}^{-1} (\boldsymbol{\mu}_{me} - \boldsymbol{\mu}_{nig}) \\ - \frac{1}{2}\log|\mathbf{\Sigma}_n| - \frac{D}{2}(1 + \log 2\pi + \frac{1}{Tw_m + r})\Big],\end{aligned} \tag{3}$$

where T is the total number of frames for training λ_e, D is features dimension and r is a relevance factor that controls compromise between UBM statistics and adaptation data in GMM adaptation [9]. The value $r = 0$ corresponds to barely standing on adaptation data.

2.3 Proposed method

In this work, we train the UBM (λ_{UBM}) using digitally mixed speech signals at different SSR levels formed by different speakers. Moreover, we train each target speaker i, the set of gain-dependent models λ_{ig} that are adapted from the UBM based on ith speaker speech files corrupted by other speakers signal at SSR level g. Using SSR-based speaker models, the system captures speaker-dependent information when it is contaminated by other speakers data. This is similar to the idea of having an SSR-based bias in GMM parameters in [4], however, it has the major difference that we build separate GMMs for each SSR level based on the UBM. It enables the system to function independent of the SSR level.

For a feature vector extracted from a speech segment at time instance t, and denoted by $\mathbf{x}_t$, frame level score for speaker i is defined as,

$$s_{it} = \frac{1}{G}\sum_{g=1}^{G} \log[p(\mathbf{x}_t|\lambda_{ig})] - \log[p(\mathbf{x}_t|\lambda_{UBM})], \tag{4}$$

We average over all SSR levels to be independent of the underlying SSR in the given signal and normalize all speakers scores at time instance t with the corresponding UBM score. To emphasize dominant speaker score in a frame, the score in (4) is further normalized by $s'_{it} = s_{it}/\sigma_t$, where σ_t is standard deviation of all

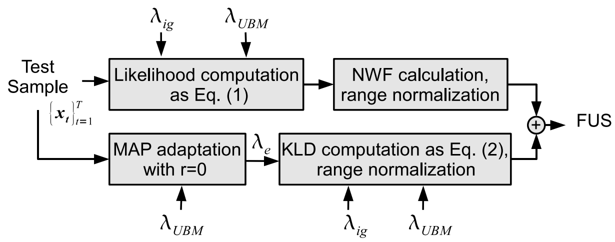

Figure 1. Proposed SID module is a combination of frame level likelihood score and model level distance: $\mathrm{FUS} = 0.54\mathrm{NWF} + 0.46\mathrm{KLD}$.

speakers scores for the frame t. To sum up, we consider five different scores for a speaker:

NWF: *number of winning frames*, where speaker i is the most probable speaker in that frame, $NWF_i = \sum_t \varphi(s'_{it})$ where $\varphi(s'_{it}) = 1$ for $i = \arg\max_j s'_{jt}$ and 0 otherwise.

NCF: *number of confident frames* for speaker i where s'_{it} is above threshold α: $NCF_i = \sum_t \psi(s'_{it})$ where $\psi(s'_{it}) = 1$ for $s'_{it} > \alpha$ and 0 otherwise.

LL: *Log-likelihood* mean for which s'_{it} is above threshold α: $LL_i = (1/NCF_i) \sum_t \psi(s'_{it}) s'_{it}$.

KLD: *Kullback-Leibler divergence* between λ_e and a set of models λ_{ig}, computed using (2).

ACE: *approximate cross entropy* between λ_e and a set of models λ_{ig}, computed using (3).

As it is common in speaker recognition, to enable using benefits from different recognizers, we considered the fusion of the scores. We used an approximate brute-force search to find the optimal weights for score fusion. It should be mentioned that we normalized (and reverted for KLD) the range of scores from different recognizers before fusion. A block diagram of proposed system is presented in Fig. 1.

3 Experimental setup

We evaluate the proposed SID module using the *speech separation challenge* corpus provided in [6]. The corpus is composed of 34 speakers (18 male, 16 female), with a total number of 34,000 utterances, each following a command-like structure, and all having a unique grammatical structure. Each sentence is formed by different syntaxes of command, color, letter, number and code, for instance ”bin white by A 3 please”. The test data in the corpus is composed of 500 laboratory-quality signals for each of the 34 target speakers, as well as test set consisting of mixed signals at six signal-to-signal ratio levels of $\{-9, -6, -3, 0, 3, 6\}$ dB. For each of these six test sets for two-talker signal, 600 utterances are provided, from which 221 are for same talker (ST), 200 for same gender (SG), and 179 for different gender (DG). The utterances were originally sampled at 25 kHz with a duration of 2 second.

Since we are interested in telephone-quality speech bandwidth, we downsample the signals from 25 kHz to 8 kHz. We extract features from 30 msec frames multiplied by a Hamming window. A 27-channel mel-frequency filterbank is applied on discrete Fourier transform (DFT) spectrum to extract 12-dimensional mel-frequency cepstral coefficients (MFCCs), followed by appending Δ and Δ^2 coefficients, and using an energy-based voice activity detector (VAD) for extracting the feature vectors. We digitally add the signals with an average frame-level SSR to construct the UBM and the target speakers GMMs. For each of 34 speakers, 50 random files from each speaker were mixed at SSRs levels $\{-9, -6, -3, 0, 3, 6\}$ dB with 50 random files from other speakers which gives us about 180 hour of speech for training UBM. The number of Gaussians, M, is set to 2048.

Speakers SSR-dependent GMMs, λ_{ig} , trained by mixing 100 random files from each speaker with 100 random files from other speakers yielding about 1.8 hours data for each SSR. Relevance factor was set to 16 for training speaker models, λ_{ig}, where its value was set to 0 in training test model, λ_e, because of availability of only 2 seconds of data for adaptation. We set the threshold α to 1 in frame-level scores calculation. The accuracies defined here are to identify both of the speakers existing in mixed signal as the two most probable speakers.

4 Experimental results

We first analyze the performance of speaker identification system using each of the 5 scores individually. The results shown in Table 1 indicate that NWF and KLD have the best average performance compared to the other methods. To the best of our knowledge, SID accuracy for *Iroquois* is not reported without SSR estimation included. According to the system configuration we used, and since we use the MAP adapted GMMs rather than the ML trained ones, we deem that *LL* method accuracy could be a simulation of *Iroquois* SID accuracy without SSR estimation. Surprisingly, compared to *LL* score, our proposed method, *NWF*, is more accurate. It is observed that, the number of frames above the confidence level, NCF is more important than their mean value, *LL*. On the other hand, the model based approach, *ACE*, works equally well as the frame-level method but it is more complex and has slightly worse accuracy than *KLD*.

Score fusion was then done by using two most suc-

Table 1. Speaker Identification accuracy for different systems (percentage of utterances with both speakers in the 2-best list output). FUS is proposed system composed of 0.54*NWF* $+$ 0.46*KLD* and *IRO* stands for Iroquois

SSR (dB)	-9	-6	-3	0	3	6	Ave
NWF	81	90	94	95	92	88	90
NCF	75	88	93	94	92	86	89
LL	74	84	90	91	87	82	85
KLD	79	89	92	93	91	87	88
ACE	79	87	92	92	89	84	87
FUS	92	93	96	97	93	87	93
IRO [4]	96	98	98	99	99	98	98

Table 2. Speaker Identification accuracy for proposed *FUS* system (percentage of utterances with both speakers in the 3-best list output) ST, Same Talker, SG, Same Gender and DG, Different Gender).

SSR	ST	SG	DG	Ave
-9 dB	100	93	83	92
-6 dB	100	97	94	97
-3 dB	100	100	98	99
0 dB	100	98	99	99
3 dB	100	97	93	97
6 dB	100	94	91	95
Ave	100	97	93	97

cessful methods: $\mathrm{FUS}_i = 0.54\mathrm{NWF}_i + 0.46\mathrm{KLD}_i$. The fusion weights were optimized on development set consisting of 300 mixed signals for each SSR level. We found that, for the fusion system, in all of the experiments, one of the speakers in the mixed signal is *always* identified. The accuracy of the proposed system (FUS) for listing two target speakers in 3-best list is shown in Table 2. This accuracy suggests to use proposed SID module as a concise ”short-list” generator for the SSR estimation in *Iroquois* to reduce complexity. To understand the system performance better, we look for combinations of speakers that are identified in any given SSR. Surprisingly, in 68% of cases both speakers are correctly identified in the mixed signal at all SSR levels, and in 80% of experiments possibly only for one SSR we cannot identify both speakers but one of them. From the results, it is observed that mixed signals with different genders (DG) are more problematic than the same gender, which there are almost no significant difference in identification accuracy between males and females.

5 Conclusion

A new method for speaker identification in co-channel scenario was introduced based on the existing approaches in speaker verification and compared the accuracy to *Iroquois* approach. From the simulation results conducted on speech separation challenge database, we observed that the proposed simple SID module performs well in listing two target speakers as three most probable speakers without any requirement on the estimates of the SSR level. As a future work, since we already got satisfactory results on 8 KHz speech, we plan to examine the proposed algorithm on telephony quality spontaneous speech and more realistically when signals are not synthetically mixed.

References

[1] T. Kinnunen and H. Li. An overview of text-independent speaker recognition: from features to supervectors. *Elsevier speech communication*, 52(1):12–40, January.

[2] D. A. van Leeuwen, A. F. Martin, M. A. Przybocki, and J. S. Bouten. NIST and NFI-TNO evaluations of automatic speaker recognition. *Elsevier Comp. Speech and Lang.*, 20(3):128–158, 2006.

[3] D. Morgan, E. George, L. Lee, and S. Kay. Co-channel speaker separation. In *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, pages 828–831, 1995.

[4] J. Hershey, S. Rennie, P. Olsen, and T. Kristjansson. Super-human multi-talker speech recognition: A graphical modeling approach. *Elsevier Comp. Speech and Lang.*, 24(1):45–66, Jan 2010.

[5] P. Mowlaee, R. Saiedi, Z.-H. Tan, M. Christensen, P. Fränti, and S. Jensen. Joint single-channel speech separation and speaker identification.

[6] M. Cooke, J. Hershey, and S. Rennie. Monaural speech separation and recognition challenge. *Elsevier Comp. Speech and Lang.*, 24(1):1–15, 2010.

[7] R. Weiss and D. Ellis. Speech separation using speaker-adapted eigenvoice speech models. *Elsevier Comp. Speech and Lang.*, 24(1):16–29, 2010.

[8] P. Li, Y. Guan, S. Wang, B. Xu, and W. Liu. Monaural speech separation based on MAXVQ and CASA for robust speech recognition. *Elsevier Comp. Speech and Lang.*, 24(1):30–44, Jan. 2010.

[9] D. Reynolds, T. Quatieri, and R. Dunn. Speaker verification using adapted gaussian mixture models. *Digital Signal Processing*, 10(1-3):19–41, Jan 2000.

[10] W. Campbell, D. Sturim, and D. Reynolds. Support vector machines using GMM supervectors for speaker verification. *Signal Processing Letters, IEEE*, 13(5):308–311, May 2006.

[11] H. Aronowitz and D. Burshtein. Efficient speaker recognition using approximated cross entropy (ACE). *Audio, Speech, and Language Processing, IEEE Transactions on*, 15(7):2033–2043, Sept. 2007.