Improving Monaural Speaker Identification by Double-Talk Detection

Saeidi, Rahim ; Mowlaee, Pejman; Kinnunen, Tomi ; Tan, Zheng-Hua; Christensen, Mads Græsbøll; Jensen, Søren Holdt; Fränti, Pasi

Published in:
Proceedings of the International Conference on Spoken Language Processing

Publication date:
2010

Document Version
Accepted author manuscript, peer reviewed version

Improving Monaural Speaker Identification by Double-Talk Detection

Rahim Saeidi[1], *Pejman Mowlaee*[2], *Tomi Kinnunen*[1],
Zheng-Hua Tan[2], *Mads Græsbøll Christensen*[3], *Søren Holdt Jensen*[2], *Pasi Fränti*[1]
[1]School of Computing University of Eastern Finland, Joensuu, Finland
[2]Dept. of Electronic Systems, and [3]Architecture Design and Media Technology,
Aalborg University, Denmark
{rahim,tkinnu,franti}@cs.joensuu.fi {pmb,zt,shj}@es.aau.dk, mgc@imi.aau.dk

Abstract

This paper describes a novel approach to improve monoaural speaker identification where two speakers are present in a single-microphone recording. The goal is to identify both of the underlying speakers in the given mixture. The proposed approach is composed of a double-talk detector (DTD) as a pre-processor and speaker identification back-end. We demonstrate that including the double-talk detector improves the speaker identification accuracy. Experiments on GRID corpus show that including the DTD improves average recognition accuracy from 96.53% to 97.43%.

Index Terms: speaker identification, double-talk detection, single-channel, Gaussian mixture models.

1. Introduction

Speaker recognition systems have evolved to reach high accuracy on clean speech signals [1]. However, speaker recognition under adverse conditions remains a challenging problem. Depending on the noise type and the way that it affects the speech signal, the more complicated methods are required to handle speaker recognition task. One of the most challenging cases are speech signals mixed with other speech signals known as monaural speech. This happens in such applications as *single-channel speech separation* [2] where accurate speaker identification is crucial for the entire system. Here we consider the task of identifying both of the speakers' identity in a given speech mixture of two speakers. Current approaches for handling this task are combined with speech separation where we cannot say exactly there is a stand-alone speaker identification system for monaural speech [3]. We have recently independently proposed methods for both speaker identification (SID) [4] and speaker-dependent double-talk detection (DTD) [5] for speech signals mixture. Our proposed method [4] does not depend on speech separation but it works directly on monaural signal without any prior information about mixing scenario of the two speech signals. In this work we improve the SID accuracy by introducing external information of mixed frames and single-talk frames provided by enhanced version of proposed DTD module in [5]. A block diagram of the proposed system is shown in Fig. 1.

Majority of the current single-channel speech separation systems use *a priori* knowledge of speaker identities [6] which is both impractical and restrictive regarding real applications. A joint system composed of speaker identification and speech separation blocks was proposed in [7] for relaxing the need for *a priori* speaker identities. The proposed system [7] improved the overall perceived speech quality of the separated output signals compared to speaker-independent and the observed speech mixture. To make speaker identification system more efficient, in this work, we introduce gender-dependent DTD and apply it to monaural SID.

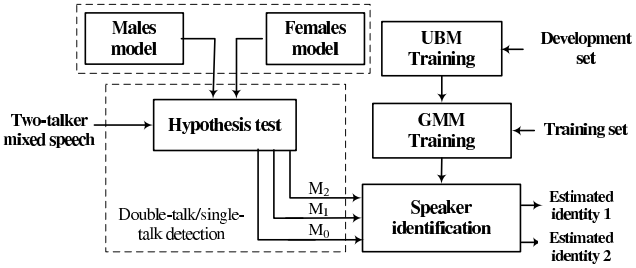

Figure 1: *The proposed method is composed of a double-talk detector followed by SSR-independent speaker identification.*

State-of-the-art single channel speech separation (SCSS) systems use a model-based SID module known as *Iroquois* [3] to identify the speakers in monaural speech. Recognition accuracies as high as 98% and 99% were reported on GRID corpus for *Iroquois* in [2,8] for locating the target speakers in short-lists of top-2 and top-3 most probable speakers respectively. In the *Iroquois* system, a short-list of the most likely speakers is produced based on frames dominated by one speaker. This short-list is then passed to a *max-based EM algorithm* for estimating both the signal-to-signal ratio (SSR) and the identities of the two speakers using exhaustive search on codebooks created for speech synthesis [3]. Based on the sizes of the short-list and code-books, this search can be time consuming. It is important to notice that if we wish to apply *Iroquois* system on a conversational mixed speech, it also requires a reliable speech separation system to produce meaningful results. Independent performance of our proposed method could be considered as a bonus in this situation. In view of this problem, the proposed system could also be used as a pre-processor for *Iroquois* system to reduce the search time.

The new contributions in this study are summarized as follows. We include a sophisticated MAP-based double talk detector (DTD) to our recent recognition system [4]. The double-talk detector was earlier introduced for monaural speech assuming known speaker identities [5]. In this paper, we adopt the method to monaural speaker identification, by using gender-dependent models to enable speaker-*independent* processing. The DTD module is utilized in the identification system so that the mixed-

The work of R. Saeidi was supported by a scholarship from the Finnish Foundation for Technology Promotion (TES). The work of P. Mowlaee is supported by the Marie Curie EST-SIGNAL Fellowship (http://est-signal.i3s.unice.fr), contract no. MEST-CT-2005-021175. The work of T. Kinnunen was supported by the Academy of Finland (project no 132129). The speaker recognition experiments were performed using computing resources from CSC (http://www.csc.fi/english) under the project no uef4836.

signal recognition score is enhanced by using "bonus" scores obtained from the more reliable single-talk regions of the mixed signal.

2. Double-Talk Detection System

In [5], a method for detecting single-talk and double-talk regions from a given speech mixture was proposed. The method was based on multiple hypothesis testing. In this work, we briefly describe the method and apply it for improving speaker identification performance. Consider monaural speech signal with N samples $\mathbf{y} = [y(0), \ldots, y(N-1)]$ composed of J speaker signals as $\mathbf{y} = \mathbf{s}_1 + \mathbf{s}_2 + \cdots + \mathbf{s}_j + \mathbf{e}$, where $j \in [1, J]$ indicates the number of signals in the mixed signal, $\mathbf{s}_j = [s_j(0), \cdots, s_j(N-1)]$ is the jth signal and $\mathbf{e}$ is the noise signal incorporated in the model. In the following, we focus on $J = 2$, that is, a mixture of two speakers.

Assume that we have K candidate models denoted by $\mathbf{M}_k$, for describing $\mathbf{y}$. The double-talk detection addresses the following problem: given the mixed signal, select the model which has the the maximum *a posteriori* (MAP) probability. We consider four models for $\mathbf{y}$ as: $\mathbf{M}_0$: $\mathbf{y} = \mathbf{e}$, $\mathbf{M}_1$:$\mathbf{y} = \hat{\mathbf{s}}_1(\{\boldsymbol{\theta}_1\})$, $\mathbf{M}_2$:$\mathbf{y} = \hat{\mathbf{s}}_2(\{\boldsymbol{\theta}_2\})$, $\mathbf{M}_3$:$\mathbf{y} = \hat{\mathbf{s}}^{(J)}(\{\boldsymbol{\theta}_1, \boldsymbol{\theta}_2\})$. Here $\hat{\mathbf{s}}_i(\{\boldsymbol{\theta}_i\})$ indicates the ith signal modeled by the parameter set $\{\boldsymbol{\theta}_i\}$ and $\hat{\mathbf{s}}^{(J)}(\{\boldsymbol{\theta}_1, \boldsymbol{\theta}_2\}) = \sum_{j=1}^{J} \hat{\mathbf{s}}_j(\{\boldsymbol{\theta}_j\})$ is the estimated mixed signal by model $\mathbf{M}_3$. Let $g_k(\mathbf{y}, \mathbf{e}, \boldsymbol{\theta}_k)$ be a generic form for class $\mathbf{M}_k$ where $k \in Z_K = \{0, 1, 2, 3\}$. Here, $\boldsymbol{\theta}_k$ is a vector composed of model parameters in a parameter space $\boldsymbol{\theta}_k \in \mathbb{R}^{m_k}$ and m_k is the length of the parameter vector $\boldsymbol{\theta}_k$. Let $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ be the vectors for model parameters for speaker one and two, respectively. Following the model selection approach in [9], we here adopt a MAP criterion for multiple-hypothesis testing to determine double-talk/single-talk regions in segments of a mixed signal. To this end, we need to evaluate the posterior probabilities of $\mathbf{M}_k$ with $k \in Z_K$. The MAP estimate of the most likely hypothesis is,

$$\hat{\mathbf{M}}_k \quad = \quad \arg \max_{\mathbf{M}_k : k \in Z_K} \left\{ \frac{p(\mathbf{y}|\mathbf{M}_k) p(\mathbf{M}_k)}{p(\mathbf{y})} \right\}, \qquad (1)$$

where $p(\mathbf{y})$ denotes the marginal density of the observed signal and $p(\mathbf{M}_k)$ is the *a priori* probability of the model $\mathbf{M}_k$. Assuming that the underlying models are equiprobable, $P(\mathbf{M}_k) = \frac{1}{K}$, dropping K and $p(\mathbf{y})$ since they are independent of M_k, the model selection rule becomes

$$\hat{\mathbf{M}}_k = \arg \max_{\mathbf{M}_k : k \in Z_K} \left\{ \int_{\boldsymbol{\theta}_k} p(\mathbf{y}|\boldsymbol{\theta}_k, \mathbf{M}_k) p(\boldsymbol{\theta}_k|\mathbf{M}_k) d\boldsymbol{\theta}_k \right\} \qquad (2)$$

where $\hat{\mathbf{M}}_k$ is the best model which achieves the MAP probability and the argument in (2) is basically $p(\mathbf{y}|\mathbf{M}_k)$. The integral in (2) is a complicated nonlinear minimization problem which can be solved by, for instance, *Laplace*s method for integration. According to [9], instead of numerical integration for the evaluation of marginal density in (2), we employ *asymptotically* MAP criterion as

$$\hat{\mathbf{M}}_k = \arg \min_{\mathbf{M}_k : k \in Z_K} \left\{ -L(\hat{\boldsymbol{\theta}}_k) + p_c \right\}, \qquad (3)$$

where p_c is the penalty of the MAP criterion and $-L(\hat{\boldsymbol{\theta}}_k)$ is log-likelihood term, given $\mathbf{M}_k$. Let $\hat{\boldsymbol{\theta}}_k$ be our feature parameters for the kth model, $\mathbf{M}_k$. As our signal modeling, to find $\hat{\boldsymbol{\theta}}_k$, we use *sinusoidal modeling* described in [7] which is based on selecting one peak per frequency band. Let $\mathbf{e}_i$ be the residual signal due to the sinusoidal modeling error in the ith band indicated by $\mathbf{e}_i = \mathbf{y}_i - \mathbf{s}_{1,i}(\boldsymbol{\theta}_1)$, where σ_i denotes the variance of the error signal in the ith band, $\mathbf{e}_i$, due to the modeling error and $\boldsymbol{\theta}_1$ is the parameter vector of length $3 \times L$ for the first speaker composed of sinusoidal parameters, L being the model order of sinusoids. Given the independence assumption in the frequency bands in subband decomposition, likelihood function for all Q bands is

$$p(\mathbf{e}|\sigma^2) = \frac{1}{(2\pi)^{\frac{N}{2}} \prod_{i=1}^{Q} \sigma_i} \exp\left(-\frac{1}{2} \sum_{i=1}^{Q} \frac{\mathbf{e}_i \mathbf{e}_i^{\mathrm{T}}}{2\sigma_i^2} \right), \qquad (4)$$

where $(\cdot)^{\mathrm{H}}$ represents the Hermitian operator. A similar expression goes for the second speaker class, $\mathbf{M}_2$ just by replacing $\mathbf{e}_i = \mathbf{y}_i - \mathbf{s}_{2,i}(\boldsymbol{\theta}_2)$ in (4), where $\boldsymbol{\theta}_2$ is the parametric vector for the second speaker.

We also include the noise model as one of the examined models by setting $g(\mathbf{y}, \mathbf{e}, \boldsymbol{\theta}) = \mathbf{e}$ and setting the number of sinusoids equal to zero ($L = 0$). We define $p(\mathbf{e}|\sigma_0^2)$ as the probability density function, with $\mathbf{e}$ considered as zero mean Gaussian noise whose noise variance is estimated by $\hat{\sigma}_0^2 = \frac{1}{N}\mathbf{y}\mathbf{y}^T$ and likelihood function given by (4).

As our last hypothesis, we are required to include the mixture model, $\mathbf{M}_3$ where the residual signal for the ith band is considered as a colored noise not fitted by $\mathbf{M}_3$ denoted by $\mathbf{e}_i = \mathbf{y} - \hat{\mathbf{s}}^{(J)}(\{\boldsymbol{\theta}_1, \boldsymbol{\theta}_2\})$. The negative log-likelihood function for mixture model $\mathbf{M}_3$ is

$$-\ln p(\mathbf{y}|\{\boldsymbol{\theta}_1, \boldsymbol{\theta}_2\}, \hat{\sigma}_i^2, \mathbf{M}_3) = \frac{N}{2} \sum_{i=1}^{Q} \ln(2\pi\hat{\sigma}_i^2) + \frac{1}{2} \sum_{i=1}^{Q} \frac{\mathbf{e}_i \mathbf{e}_i^{\mathrm{T}}}{\hat{\sigma}_i^2}. \qquad (5)$$

In order to form the MAP criterion in (3), we employ the MAP criterion [9] for sinusoids composed of amplitude and unknown frequencies and $\hat{\mathbf{M}}_k$ is obtained as

$$\hat{\mathbf{M}}_k = \arg \min_{\mathbf{M}_k \in Z_K} \left\{ \frac{N}{2} \sum_{i=1}^{Q} \ln \hat{\sigma}_i^2 + \frac{5L}{2} \ln N \right\}, \qquad (6)$$

where $\hat{\sigma}_i$ is the estimated variance for the modeling error defined for each model. For mixed class, $\hat{\mathbf{M}}_3$, as our mix model denoted by $\hat{\mathbf{s}}^{(J)}(\{\boldsymbol{\theta}_1, \boldsymbol{\theta}_2\})$, we use the minimum mean square error (MMSE) mixture estimate presented recently in [10]. According to (6), the best model, as a result, is the one which yields high log-likelihood and low model order, which is achieved according to (6) [5].

Figure 2 shows the clean signal for two speakers together with their mixture. It is observed that, the double-talk detector effectively finds the boundaries of single-talk regions. Comparing with the ground-truth, it accurately determines for each frame that which speaker(s), if any, is active. It is important to note that, for same gender or same talker scenarios DTD module degenerates into a three-class problem since it only employs one speaker model for these scenarios. Then, the double-talk detector cannot distinguish between $\mathbf{M}_1$ and $\mathbf{M}_2$, since the residual signals of these classes, are the same. The double-talk detector, however, can still identify single/double-talk regions and pass this information to the SID module.

3. Speaker Identification System

The speaker identification module is based on maximum *a posteriori* (MAP) adapted Gaussian mixture models (GMM) [11]. A speaker GMM is a weighted linear combination of M unimodal Gaussian densities where, letting λ denote a model of

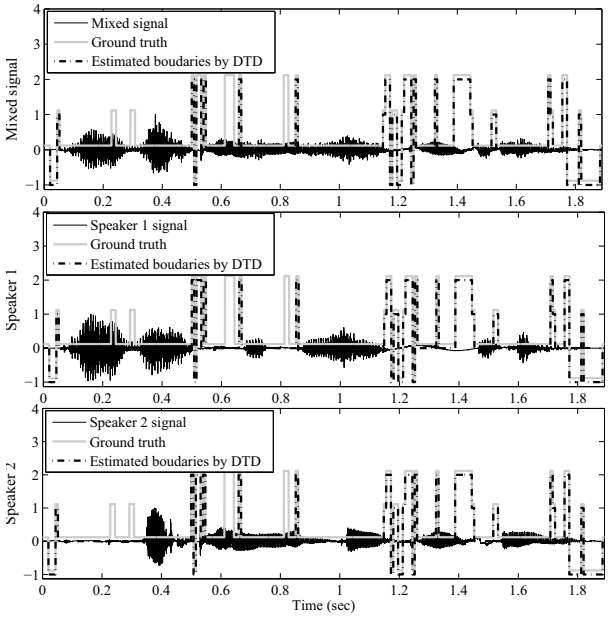

Figure 2: *Double-talk detection results for a speech mixture of a male and a female speaker mixed at 3 dB SSR. The mixed signal is composed of a male speaker 12 uttering "Lay white with e 8 again" with female speaker 11 uttering "Set green with v 3 soon". Decisions are -1 for no speech, 1 for speaker 1, 2 for speaker 2 and 0 for mixed signal regions.*

single speaker, the likelihood function is defined as,

$$\ell(\mathbf{x}) = p(\mathbf{x}|\lambda) = \sum_{m=1}^{M} w_m p_m(\mathbf{x}), \tag{7}$$

where $p_m(\mathbf{x}) = \mathcal{N}(\mathbf{x}; \mu_m, \mathbf{\Sigma}_m)$ 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 a so called *universal background model* (UBM) [11]. The UBM is a GMM trained on a pool of feature vectors extracted from as many speakers as possible and it serves as *a priori* information for feature distribution. By defining λ_{ig} as the signal-to-signal ratio dependent model for the ith speaker at SSR level g, we use frame-level likelihood and model-level approximate *Kullback-Leibler divergence* (KLD) as the similarity and distance measures respectively. For a feature vector $\mathbf{x}_t$ extracted from a speech segment at time instance t, frame level score for speaker i is defined as $s_{it} = \frac{1}{G}\sum_{g=1}^{G} s_{igt}$, where

$$s_{igt} = \log[p(\mathbf{x}_t|\lambda_{ig})] - \log[p(\mathbf{x}_t|\lambda_{UBM})]. \tag{8}$$

For speaker identification, we average over all SSR levels to make the system less dependent on the SSR level [4]. Meanwhile we normalize all speakers scores at time instance t with the corresponding UBM score. 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. Since KLD distance cannot be directly evaluated for GMMs, we use the upper bound of KLD which has successfully been applied to speaker verification [12]:

Algorithm 3.1: SPEAKER IDENTIFICATION()

Inputs:
- $\{\mathbf{x}_t\}_{t=1}^{T}$:Test sample feature vectors.
- λ_{ig} :One GMM per speaker per SSR level.
- λ_{UBM} :Background model.
- DTD Labels:$\in$ {Silent, Mixed, Spk1, Spk2}

for Mixed frames of length T_0, $\{\mathbf{x}_t, t \in$ Mixed$\}$

- **comment:** Computing FLL
- **for** $i \leftarrow 1$ **to** N
  - **for** $t \leftarrow 1$ **to** T
    - **for** $g \leftarrow 1$ **to** G
      - compute $s[i,g,t]$
    - $s[i,t] \leftarrow \frac{1}{G}\sum_{g=1}^{G} s[i,g,t]$
    - $s'[i,t] \leftarrow \arg\max_{g}\{s[i,g,t]\}$
  - $\varphi(s[i,t]) = \begin{cases} 1 & i = \arg\max_{j}\{s[j,t]\} \\ 0 & \text{otherwise.} \end{cases}$
  - $\mathrm{FLL}[i] = \sum_{t} \varphi(s[i,t])$
- $\mathrm{FLL}[i] = \frac{\mathrm{FLL}[i] - \mathrm{minFLL}}{\mathrm{maxFLL} - \mathrm{minFLL}}$
- **comment:** Computing KLD
- $\lambda_e \leftarrow$ GMMADAPT$(\{\mathbf{x}_t\}_{t=1}^{T}, \lambda_{UBM})$
- **for** $i \leftarrow 1$ **to** N
  - **for** $g \leftarrow 1$ **to** G
    - compute $\mathrm{KLD}[i,g]$
  - $\mathrm{KLD}[i] \leftarrow \frac{1}{G}\sum_{g=1}^{G} \mathrm{KLD}[i,g]$
- $\mathrm{KLD}[i] = 1 - \frac{\mathrm{KLD}[i] - \mathrm{minKLD}}{\mathrm{maxKLD} - \mathrm{minKLD}}$
- $score = 0.5 \times \mathrm{KLD} + 0.5 \times \mathrm{FLL}$

for each set of single-talk frames of length T_1 and T_2, $\{\mathbf{x}_t, t \in$ Spk1$\}$ and $\{\mathbf{x}_t, t \in$ Spk2$\}$

- **do** Compute KLD
- $idx = \arg\min\{\mathrm{KLD}\}$
- $score[idx] = score[idx] + \alpha T_1/T($ or $\alpha T_2/T)$

$$\mathrm{KLD}_{ig} = \tfrac{1}{2}\sum_{m=1}^{M} w_m(\boldsymbol{\mu}_{me} - \boldsymbol{\mu}_{mig})^{\mathrm{T}}\mathbf{\Sigma}_m^{-1} (\boldsymbol{\mu}_{me} - \boldsymbol{\mu}_{mig}). \tag{9}$$

Here g ranges over a discrete 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. To sum up, we consider two different scores for a speaker:

FLL**:** *Frame level likelihood*, where we are considering number of winning frames that speaker i is the most probable speaker in that frame for speaker identification.

KLD**:** *Kullback-Leibler divergence* between λ_e and a set of models λ_{ig}, computed using (9). We form an $N \times G$ distance matrix and average over SSR levels to raise the speaker with minimum average distance.

As commonly done in speaker recognition, to enable using benefits from different recognizers, we considered the fusion of the scores with equal weights. Similar to [4], each speaker's decision score is computed as $0.5 \times \mathrm{FLL} + 0.5 \times \mathrm{KLD}$. The frames detected by DTD module to belong to a single speaker only (1 or 2) are collected accordingly and passed to KLD score computation. Since we believe that these frames belong to only one speaker, for the speaker that gets the minimum KLD, we add a bonus score to it's decision score as

$score[idx] = score[idx] + \alpha T_1/T$ (or $\alpha T_2/T$) where idx is the identified speaker from single-talk frames. The bonus is made relative to the number of single-talk frames identified to belong to speaker 1 or 2 (T_1 or T_2) respect to total number of frames in a given test signal (T). The stressing factor α is a control parameter. Details of the SID algorithm presented as a pseudocode in Algorithm 3.1.

4. Experimental results

We evaluate the proposed system on the speech separation database known as GRID corpus [2] composed of 34,000 different utterances. The sentences were originally sampled at 25 kHz with a duration of 2 seconds each. As we usually deal with 8 kHz speech in most of speech applications, we decreased the sampling rate down to 8 kHz. The speaker models used for DTD module are split-VQ codebooks [7] composed of sinusoidal amplitude and frequencies. For training the speaker models, we used 11 bits for amplitude and 3 bits for frequency part. To train gender-dependent models, we selected 10 female and 10 male speakers each producing 35 s of speech signal. Throughout the experiments, a Hamming window of length 32 ms with frame-shift equal to 8 ms is used to segment the speech files both in the training and test phases. As our test data, we used the mixture of target and masker speakers in the test setup of [3] mixed at six SSR levels of $\{-9, -6, -3, 0, 3, 6\}$ dB. The codebook size for split-VQ was M=2048 and the sinusoidal model order was set to 50.

For speaker identification, we extract features from 30 ms frames multiplied by a Hamming window. A 27-channel mel-frequency filterbank is applied on 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 model order of the GMM is set to 2048.

The 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. For each six test sets of two-talker signal, 600 utterances were provided among which 200 were for same gender (SG), 179 for different gender (DG), and 221 for same talker (ST) where the target and masker signals are from the same speaker. To incorporate the bonus for single-talk detected frames, we used $\alpha = 5$.

Speaker identification results for the combined system presented in Table 1. Compared to the previous results without DTD [4], embedding the DTD module enhances performance. The improvement is higher on the different gender (DG) case where the gender-dependent DTD module distinguishes between single-talk areas for two speakers accordingly. Compared to the reported accuracy of 99% for the *Iroquois* system for detecting target speakers among three most probable cases [2], the proposed system achieves a comparable rate of 97.43%. Given its relatively low complexity, our proposed system could be considered as an alternative or a pre-processing block for *Iroquois* system.

Table 1: Speaker identification accuracy (% correct) where both speakers are correctly found in the top-3 list. Yes/No indicates whether the proposed DTD method is included. For the ST scenario both of the systems provide 100 % accuracy.

	SG		DG		Average	
DTD	No	Yes	No	Yes	No	Yes
SSR						
-9 dB	92.74	93.30	82.50	86.97	92.00	94.68
-6 dB	96.65	96.65	94.00	95.00	97.00	97.71
-3 dB	99.44	99.44	97.50	98.00	99.00	99.39
0 dB	98.32	98.32	99.00	98.00	99.17	99.39
3 dB	97.21	97.77	93.50	95.00	97.00	98.11
6 dB	93.85	94.41	90.50	89.50	95.00	95.63
Average	96.36	96.65	92.83	93.83	96.53	97.43

5. Conclusions

We introduced gender-dependent double talk detector for monaural speech and applied it in speaker identification task for. Speaker identification results on GRID corpus demonstrated the improvement over the system without DTD. Overall speaker identification performance is close to the results of the *Iroquois* system using computationally simple approach.

6. References

[1] T. Kinnunen and H. Li, "An overview of text-independent speaker recognition: from features to supervectors," *Elsevier speech communication*, vol. 52, no. 1, pp. 12–40, January 2010.

[2] M. Cooke, J.R. Hershey, and S.J. Rennie, "An audio-visual corpus for speech perception and automatic speech recognition," *Elsevier Computer Speech and Language*, vol. 24, no. 1, pp. 1–15, 2010.

[3] J.R. Hershey, S.J. Rennie, P.A. Olsen, and T.T. Kristjansson, "Super-human multi-talker speech recognition: A graphical modeling approach," *Elsevier Computer Speech and Language*, vol. 24, no. 1, pp. 45–66, Jan 2010.

[4] R. Saeidi, P. Mowlaee, T. Kinnunen, Z.-H Tan, M.G. Christensen, S.H. Jensen, and P. Fränti, "Signal-to-signal ratio independent speaker identification for co-channel speech signals," in *Proc. IEEE Int. Conf. Pattern Recognition*, Aug. 2010.

[5] P. Mowlaee, M.G. Christensen, Z.-H. Tan, and S.H. Jensen, "A MAP criterion for detecting the number of speakers at frame level in model-based single-channel speech separation," in *Proc. Asilomar Conf. Signals, Systems, and Computers*, Aug. 2010.

[6] S. Roweis, "Factorial models and refiltering for speech separation and denoising," in *EUROSPEECH*, 2003, pp. 1009–1012.

[7] P. Mowlaee, R. Saeidi, Z.-H. Tan, M.G. Christensen, P. Fränti, and S.H. Jensen, "Joint single-channel speech separation and speaker identification," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, March 2010, pp. 4430–4433.

[8] T. Kristjansson, J. Hershey, P. Olsen, S. Rennie, and R. Gopinath, "Super-human multi-talker speech recognition: The IBM 2006 speech separation challenge system," in *Proc. Interspeech*, 2006, pp. 97–100.

[9] P.M. Djuric, "Asymptotic MAP criteria for model selection," *Signal Processing, IEEE Transactions on*, vol. 46, no. 10, pp. 2726–2735, Oct 1998.

[10] P. Mowlaee, M.G. Christensen, and S.H. Jensen, "Sinusoidal masks for single channel speech separation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, March 2010, pp. 4262–4266.

[11] D.A. Reynolds, T.F. Quatieri, and R.B. Dunn, "Speaker verification using adapted gaussian mixture models," *Digital Signal Processing*, vol. 10, no. 1-3, pp. 19–41, Jan 2000.

[12] W.M. Campbell, D.E. Sturim, and D.A. Reynolds, "Support vector machines using GMM supervectors for speaker verification," *Sig. Proc. Letters, IEEE*, vol. 13, no. 5, pp. 308–311, May 2006.