Speech Communication 87 (2017) 40–48

Contents lists available at ScienceDirect

Speech Communication

journal homepage: www.elsevier.com/locate/specom

Influence of binary mask estimation errors on robust speaker identification

Tobias May*

Hearing Systems Group, Department of Electrical Engineering, Technical University of Denmark, Kgs. Lyngby DK - 2800, Denmark

ARTICLE INFO

Article history:
Received 8 March 2016
Revised 5 December 2016
Accepted 9 December 2016
Available online 16 December 2016

Keywords:
Speaker identification
Missing data
Ideal binary mask
Estimated binary mask
Bounded marginalization
Full marginalization
Direct masking

ABSTRACT

Missing-data strategies have been developed to improve the noise-robustness of automatic speech recognition systems in adverse acoustic conditions. This is achieved by classifying time-frequency (T-F) units into reliable and unreliable components, as indicated by a so-called binary mask. Different approaches have been proposed to handle unreliable feature components, each with distinct advantages. The direct masking (DM) approach attenuates unreliable T-F units in the spectral domain, which allows the extraction of conventionally used mel-frequency cepstral coefficients (MFCCs). Instead of attenuating unreliable components in the feature extraction front-end, full marginalization (FM) discards unreliable feature components in the classification back-end. Finally, bounded marginalization (BM) can be used to combine the evidence from both reliable and unreliable feature components during classification. Since each of these approaches utilizes the knowledge about reliable and unreliable feature components in a different way, they will respond differently to estimation errors in the binary mask. The goal of this study was to identify the most effective strategy to exploit knowledge about reliable and unreliable feature components in the context of automatic speaker identification (SID). A systematic evaluation under ideal and non-ideal conditions demonstrated that the robustness to errors in the binary mask varied substantially across the different missing-data strategies. Moreover, full and bounded marginalization showed complementary performances in stationary and non-stationary background noises and were subsequently combined using a simple score fusion. This approach consistently outperformed individual SID systems in all considered experimental conditions.

1. Introduction

The automatic identification of speakers in adverse scenarios is an important building block for many applications, including access control, authentication, personalization of communication services and forensic applications (Campbell, 1997; Campbell et al., 2009). Conventional speaker identification (SID) systems employ mel-frequency cepstral coefficients (MFCCs) in combination with Gaussian mixture model (GMM) classifiers (Reynolds and Rose, 1995) and universal background model (UBM) adaptation (Reynolds et al., 2000). To increase the robustness of SID systems against changes in the acoustic environment between the training and testing stage, feature normalization strategies are usually applied. The most commonly used normalization strategies are file-based or adaptive cepstral mean and variance normalization (Viikki and Laurila, 1998) as well as histogram equalization (HEQ) techniques (de la Torre et al., 2005).

In contrast to conventional approaches, the missing-data technique classifies the time-frequency (T-F) representation of noisy speech into target-dominated (reliable) and interference-dominated (unreliable) T-F units (Cooke et al., 2001). This binary decision can, for example, be accomplished by the so-called ideal binary mask (IBM), which assumes *a priori* knowledge about the energy of the target and the interfering noise (Wang, 2005). The recognition of speaker identities is subsequently performed using only those T-F units that are believed to be reliable. Many studies have demonstrated that the missing-data technique improves the robustness of automatic speech recognition (Vizinho et al., 1999; Cooke et al., 2001; Ris and Dupont, 2001) and SID systems (Drygajlo and El-Maliki, 1998; Jančovič and Köküer, 2006; Shao and Wang, 2006; May et al., 2012a; 2012b; Zhao et al., 2014) in noisy environments. If, instead of a binary classification, an estimation of the feature uncertainty is available, this information can be ex-

* Corresponding author.
E-mail address: tobmay@elektro.dtu.dk

http://dx.doi.org/10.1016/j.specom.2016.12.002

ploited by an uncertainty decoder for improved robustness (Deng et al., 2005; Shao et al., 2007; Ozerov et al., 2013; Yu et al., 2014). For a comprehensive overview the reader is referred to Kolossa and Haeb-Umbach (2011).

Assuming that a binary decision about the feature reliability is available, two different approaches exist to incorporate the knowledge about reliable and unreliable T-F units in the classification back-end. The first - full marginalization (FM) - ignores all unreliable T-F units and estimates the likelihood of a particular speaker identity based on reliable T-F units only. Despite being dominated by noise, unreliable T-F units can be used to provide *counter-evidence* based on the concept of energetic masking (Cooke et al., 2001). Consequently, the second approach - bounded marginalization (BM) - additionally exploits knowledge about unreliable T-F units for improved identification performance by effectively penalizing less energetic speaker models.

One of the most critical factors that limits the performance of missing-data strategies is the accuracy of the binary mask. When *a priori* knowledge in form of the IBM is available, the BM approach is very effective and produces SID scores above 95%, even at negative signal-to-noise ratios (SNRs) (May et al., 2012b). However, when estimated binary masks (EBMs) are used instead, the achievable performance of BM decreases due to misclassified T-F units in the binary mask, most noticeably in highly non-stationary background noises (May et al., 2012b). In addition, BM treats reliable and unreliable feature components differently during classification and, thus, requires a correspondence between the individual T-F units in the binary mask and the feature components used for recognition. Therefore, the BM approach is limited to spectral features, such as filter-bank energy (FBE) features, which are known to be less powerful than their decorrelated counterparts, such as MFCC features, which operate in the cepstral domain.

The performance of FBE features can be improved by applying a derivative finite impulse response (FIR) filter across frequency channels, which produces decorrelated filter-bank energy (DFBE) features (Nadeu et al., 1995; 2001). These DFBE features achieve similar SID performance compared to MFCC features, with the advantage that frequency-specific distortions due to background noise remain local and can still be associated with a restricted set of T-F units. While it is not beneficial to exploit information about unreliable T-F units after the FIR-based decorrelation stage due to the wide feature bounds (Barker, 2012), DFBE features can be combined with FM, where unreliable feature components are ignored during classification.

Alternatively, the knowledge about reliable and unreliable T-F units can be exploited in the feature extraction front-end. Specifically, the spectral representation of noisy speech can be enhanced by a binary or continuous gain function (El-Solh et al., 2007; Sadjadi and Hansen, 2010; Jensen and Hendriks, 2012; Godin et al., 2013). This allows the use of conventional decorrelation stages, such as the discrete cosine transform (DCT), to convert the modified FBE features to MFCC features. When assuming *a priori* knowledge about the noise power, it was shown that an *ideal* noise reduction scheme based on a continuous gain function achieved the same SID performance compared to IBM-based BM (May et al., 2012b). Likewise, the direct masking (DM) approach applies the IBM directly to the FBE features, and was reported to produce similar automatic speech recognition performance compared to the BM strategy (Hartmann et al., 2013). However, estimation errors in the gain function can distort the resulting feature vector, which may limit the effectiveness under realistic conditions.

Each of the aforementioned missing-data methods has distinct advantages and applies the binary mask in a different way. As a result, mask estimation errors will have different consequences on the achievable speaker identification performance. Nevertheless, the influence of binary mask errors on the different missing-data strategies has not yet been systematically investigated. The majority of studies obtained an estimation of the binary mask by predicting the local SNR in individual T-F units (Drygajlo and El-Maliki, 1998; Renevey and Drygajlo, 2000; Cooke et al., 2001; Ris and Dupont, 2001; May et al., 2012b). Recently, supervised learning approaches have reported increased mask estimation accuracies by exploiting *a prior* knowledge about the distribution of acoustic features observed during an initial training phase (Seltzer et al., 2004; May and Dau, 2013; 2014; Zhao et al., 2014). However, the general advantage of missing-data strategies is that speaker models are trained with clean speech only, and no prior knowledge about the acoustic environment or about the interfering noise is required.

SID systems based on the GMM-UBM back-end are the most commonly-used systems when applying missing-data strategies (Togneri and Pullella, 2011; May et al., 2012a; 2012b; Zhao et al., 2012; 2014). Lately, the i-vector approach has received increasing attention, in particular in the field of speaker verification, by considering the inter- and intra-speaker variability of the feature vector (Dehak et al., 2011). A recent comparative study of noise reduction strategies (e.g. power spectral subtraction, Wiener filtering and log-minimum mean-square error speech enhancement) reported similar relative improvements over a MFCC baseline system for both GMM-UBM and i-vector-based SID systems (Godin et al., 2013). From this perspective, a similar benefit can be expected when the DM approach is combined with i-vector-based SID systems.

The goal of this study was to determine the effectiveness of full marginalization, bounded marginalization and direct masking in the context of closed-set SID under ideal and non-ideal conditions. To facilitate a proper comparison, a unified framework was used to optimize the criteria for deriving ideal and estimated binary masks for each missing-data system separately. In non-ideal conditions, EBMs were obtained by applying a threshold criterion to the estimated speech presence probability (SPP) in individual T-F units, which was shown to produce competitive results compared to supervised learning approaches (May and Gerkmann, 2014). A systematic comparison between ideal and non-ideal conditions was performed to analyze the robustness of the different missing-data approaches to estimation errors in the binary mask. Moreover, the SID performance was compared to a conventional noise reduction scheme. Afterwards, the noise-robustness of the most successful missing-data systems was compared to a conventional MFCC system across a wide range of acoustic conditions. Finally, a simple score fusion was tested which combined two missing-data systems with complementary advantages in stationary and non-stationary background noises. To support reproducible research, the complete Matlab code of the SID framework, which was used to produce all experimental results in the present study, is available online (May, 2016).

2. System

The SID framework shown in Fig. 1 consisted of a training and a testing stage. First, speaker models were trained with features extracted from clean speech. In the testing stage, features were derived from noisy speech and the reliability of individual feature components was estimated by means of a binary mask. The subsequent classification stage was configured to use three different missing-data strategies, namely full marginalization (FM), bounded marginalization (BM) and direct masking (DM). Depending on which type of missing-data strategy was used, different feature decorrelation and normalization stages were combined. A list of tested configurations is shown in Table 1. Each processing stage is described in detail in the following.