Neurocomputing 78 (2012) 111–120

Contents lists available at SciVerse ScienceDirect

Neurocomputing

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

Environmental robust speech and speaker recognition through multi-channel histogram equalization

Stefano Squartini *, Emanuele Principi, Rudy Rotili, Francesco Piazza

3MediaLabs, Department of Information Engineering, Università Politecnica delle Marche, Via Brecce Bianche 1, 60131, Ancona, Italy

ARTICLE INFO

Available online 25 August 2011

Keywords:
Multi-channel audio processing
Feature statistics normalization
Histogram equalization
Speech recognition
Speaker recognition

ABSTRACT

Feature statistics normalization in the cepstral domain is one of the most performing approaches for robust automaticspeech and speaker recognition in noisy acoustic scenarios: feature coefficients are normalized by using suitable linear or nonlinear transformations in order to match the noisy speech statistics to the clean speech one. Histogram equalization (HEQ) belongs to such a category of algorithms and has proved to be effective on purpose and therefore taken here as reference.

In this paper the presence of multi-channel acoustic channels is used to enhance the statistics modeling capabilities of the HEQ algorithm, by exploiting the availability of multiple noisy speech occurrences, with the aim of maximizing the effectiveness of the cepstra normalization process. Computer simulations based on the Aurora 2 database in speech and speaker recognition scenarios have shown that a significant recognition improvement with respect to the single-channel counterpart and other multi-channel techniques can be achieved confirming the effectiveness of the idea. The proposed algorithmic configuration has also been combined with the kernel estimation technique in order to further improve the speech recognition performances.

1. Introduction

Speech interfaced Human–Machine systems have been gaining an increasing interest among the scientific community and technology market. In the last decades, a great deal of research has been devoted both to extending our capacity of verbal communication with computers and also to let machines automatically identify or verify speaking people by means of their voice: more and more performing automatic speech and speaker recognition systems have been studied and developed on purpose, respectively.

Although optimum performance can be reached when the speech signal is captured close to the speaker's mouth, there are still obstacles to overcome in making reliable distant speech/speaker recognition technologies. The two major sources of degradation are additive noise and reverberation. This implies that speech enhancement techniques [1] are typically required to improve the quality of the observed signal. Different methodologies have been proposed in the literature for environment robustness in speech/speaker recognition over the past two decades. A brief analysis of related states of the art, starting from the single-channel approaches, is reported here.

For speech recognition, two main classes can be identified [2]. The first class encompasses the so called model-based techniques, which operate on the acoustic model to adapt or adjust its parameters so that the system fits better the distorted environment. The most popular techniques are multi-style training, parallel model combination (PMC) and the vector Taylor series (VTS) model adaptation. Although model-based techniques obtain excellent results, they require heavy modifications to the decoding stage, thus exhibiting a certain dependence on the recognition engine choice and, in most cases, a significative computational burden. Conversely, the second class directly enhances the speech signal before it is presented to the recognizer. The wide variety of algorithms in this class are named as feature-enhancement techniques: they typically do not present the aforementioned drawbacks characterizing the model-based category and can be further divided based on the number of channels used in the enhancing stage. Single-channel approaches encompass classical techniques operating in the frequency domain such as Wiener filtering, spectral subtraction and Ephraim and Malah [3,4], as well as techniques operating in the feature domain such as MFCC–MMSE [5] and its optimizations [6,7], and VTS based approaches [8–10]. Feature statistics normalization approaches, as cepstral mean–variance normalization (CMVN) [11], higher order cepstral moment normalization (HOCMN), histogram equalization (HEQ) [12], cepstral shape normalization [13] and parametric feature equalization [14], also play an important role in this regard.

A similar analysis for robust speaker recognition (i.e. both identification and verification) can be drawn following the study in [15]. It has to be underlined that the text dependency issue is not considered in this paper, since the focus is on robustness

* Corresponding author.
E-mail address: s.squartini@univpm.it (S. Squartini).

0925-2312/$ - see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.neucom.2011.05.035

and not on the employed recognition technique. More precisely, various feature-enhancement algorithms, operating both at the frequency and cepstral domain, have been proposed. Regarding the former domain, all aforementioned noise reduction techniques can be mentioned together with some more specific ones like the method in [16] which combines temporal and spectral pre-processing for speaker recognition under noisy, reverberant or multi-speaker environments. The latter encloses not only feature warping [17] and Gaussianization techniques [18], but also RASTA techniques [19], approaches combining multicondition model training and missing feature theory [20] and graph-theoretic compensation methods [21]. Some researchers have also recently investigated various combination of normalizing feature statistics [22]. As in the speech recognition case study, also some model-based approaches have been presented as the method in [23], where the speaker model is adaptively compensated in order to take the environment characteristics into account.

Multi-channel approaches use the benefits of the additional information carried out by the presence of multiple speech observations. In most cases the speech and noise sources are in different spatial locations, thus a multi-microphone system is theoretically able to obtain a significant gain over single-channel approaches, since it may exploit the spatial diversity. Three different categories of algorithms can be identified from this perspective: beamforming techniques, Bayesian estimators (operating at different level of the feature extraction pipeline) and feature statistics normalization.

In speech recognition scenario, beamforming techniques are employed as pre-processing stage. In [24] the delay and sum beamformer (DSB) has been successfully used coupled with a talker localization algorithm, but its performances are poor when it operates in a reverberant environment. This motivated the scientific community to develop more robust beamforming techniques, as the generalized sidelobe canceler (GSC) and the several GSC-based techniques appeared in the literature. In particular, in [25] a psychoacoustically motivated TF-GSC beamformer has been presented and its effectiveness demonstrated in a speech recognition scenario. Among the beamforming techniques, likelihood maximizing beamforming (LIMABEAM) is an hybrid model-based approach that uses information from the decoding stage to optimize a filter and sum beamformer [26].

Multi-channel Bayesian estimators in frequency domain have been proposed in [27] where both minimum mean square error (MMSE) and maximum a posteriori (MAP) criteria were developed. The feature domain counterpart of the previous algorithms has been presented in [28].

In the speaker recognition case study, the employment of multi-channel audio information has been typically oriented to jointly compensate the mismatch introduced by noise and reverberation, both in simulated and real scenarios. Beamforming techniques [29,30] have been mainly proposed on purpose, but some recent contributions aimed at combining the multi-microphone information and suitably using it in the training phase [31] can also be registered.

In this paper the focus is on multi-channel feature statistics normalization, and on evaluating its effectiveness when applied within the feature front-end of speech/speaker recognition systems in far-field acoustic scenarios. A preliminary investigation has been already done in the speech recognition case study in a recent work of the authors [32]. The basic concept behind the approach consists in exploiting the presence of multiple audio channels to better estimate the input signal statistics and so making the equalization capability of feature normalization process more effective. Several computer simulations have been carried out, taking the HEQ approach as reference, to show the benefits brought by the advanced idea w.r.t. what appeared in the related literature so far.

The paper outline is as follows. Section 2 describes the feature extraction pipeline and the adopted mathematical model. Section 3 introduces the multi-channel HEQ concept and proposes various algorithmic architectures for its implementation. Section 4 presents and discuss speech/speaker recognition results in a comparative fashion, taking other existing single-channel and multi-channel approaches as references. Finally, Section 5 draws conclusions and proposes future developments.

2. Feature extraction front-end

In this section the feature extraction algorithmic pipeline is described, in order to better understand the multi-channel acoustic scenario under study and the signals processed by the algorithms addressed in our work. The pipeline extracts mel-frequency cepstral coefficients (MFCC) [33], and their first and second derivatives [34,35].

Let us consider M noisy signals $y_i(t)$, M clean speech signals $x_i(t)$ and M uncorrelated noise signals $n_i(t)$, $i \in \{1,\ldots,M\}$, where t is a discrete-time index. The i-th microphone signal is given by

$$y_i(t) = x_i(t) + n_i(t). \tag{1}$$

In general, the signal $x_i(t)$ is the convolution between the speech source and the i-th room impulse response. In our case study the far-field acoustic model [27] is considered, which assumes equal amplitude and angle-dependent time difference of arrival:

$$x_i(t) = x(t-\tau_i(\theta_x)), \quad \tau_i = d\,\sin(\theta_x/c), \tag{2}$$

where τ_i is the i-th delay, d is the distance between the source and the microphone array, θ_x is the angle of arrival and c is the speed of sound.

According to Fig. 1, each input signal $y_i(t)$ is firstly pre-emphasized and windowed with a Hamming window. Then, the FFT of the signal is computed and the square of the magnitude is filtered with a bank of triangular filters equally spaced in the mel-scale. After that the energy of each band is computed and transformed with a logarithm operation. Finally, the discrete cosine transform (DCT) stage yields the static MFCC coefficients, and the $\Delta/\Delta\Delta$ stage compute the first and second derivatives.

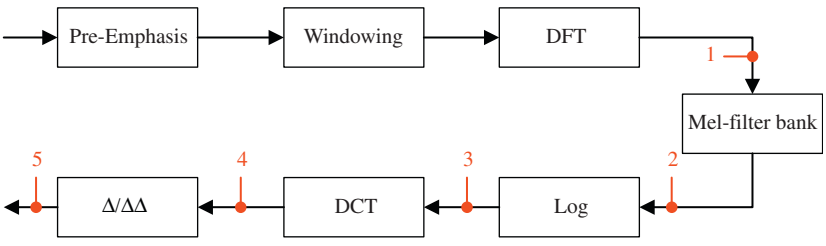

Fig. 1. Feature extraction pipeline.