Available online at www.sciencedirect.com
ScienceDirect
Computer Speech & Language xxx (2018) xxx-xxx
www.elsevier.com/locate/csl

Vocal effort compensation for MFCC feature extraction in a shouted versus normal speaker recognition task☆

Q1 Emma Jokinen[a,*], Rahim Saeidi[a], Tomi Kinnunen[b], Paavo Alku[a]

[a] Department of Signal Processing and Acoustics, Aalto University, PO Box 13000, Aalto FI-00076, Finland
[b] School of Computing, University of Eastern Finland, PO Box 111, Joensuu FI-80101, Finland

Received 1 November 2017; received in revised form 10 April 2018; accepted 6 June 2018
Available online xxx

Abstract

In shouting, speakers use increased vocal effort to convey spoken messages over distance or above environmental noise. For automatic speaker recognition systems trained using normal speech, shouting causes a severe vocal effort mismatch between the enrollment and test hence reducing the recognition performance. In this study, two compensation methods are proposed to tackle the mismatch in a shouted versus normal speaker recognition task. These techniques are applied in the feature extraction stage of a speaker recognition system to modify the spectral envelopes of shouts to be closer to those in normal speech. The techniques modify the all-pole power spectrum of the MFCC computation chain with shouted-to-normal compensation filtering that is obtained using a GMM-based statistical mapping. In an evaluation using the state-of-the-art i-vector based recognition system, the proposed techniques provided considerable improvements in identification rates compared to the case when shouted speech spectra were not processed.

Keywords: Speaker recognition; Vocal effort mismatch; Shouted speech

1. Introduction

Human speech contains a great deal of intrinsic variability, such as changes in fundamental frequency (*F*0) and intonation, different styles of speaking and phonation and different levels of vocal effort. Speaking style modifications, such as the Lombard effect which takes place when speaking in noisy conditions, are naturally used by talkers to make speech more intelligible to human listeners (Summers et al., 1988). The performance of data-driven systems, however, typically suffers when such changes in speaking style occur (Junqua, 1993). The loss in performance is due to the mismatch between the system's training conditions and its testing conditions. In this study, the focus is on severe vocal effort mismatch between the normal speaking mode and shouting. This mismatch condition is studied in automatic speaker recognition where the system has been trained using normal speech but is tested with shouted Q2 speech.

☆ This paper has been recommended for acceptance by Roger Moore.
* Corresponding author.
E-mail address: ejjokine@gmail.com (E. Jokinen), tkinnu@cs.uef.fi (T. Kinnunen), paavo.alku@aalto.fi (P. Alku).

https://doi.org/10.1016/j.csl.2018.06.002
0885-2308/

Shouting is used in situations where a message needs to be conveyed urgently over a distance or in a noisy situation. Differently from Lombard speech produced also in noisy conditions, shouted speech shows reduced intelligibility for human listeners compared to speech produced in the normal speaking mode (Pickett, 1956). While the vocal effort of Lombard speech rises to some extent from that of normal speech, the corresponding change is much more prominent when changing from normal speech to shouting. In addition to an overall sound level increase (Rostolland, 1982), also a reduction in spectral tilt (Zhang and Hansen, 2007), movement of formant frequencies (Zelinka et al., 2012; Traunmüller and Eriksson, 2000), increase in $F0$ (Rostolland, 1982) and changes in vowel and consonant durations (Rostolland, 1982; Traunmüller and Eriksson, 2000) occur when speakers change their normal speaking style to shouting.

Several studies have indicated that severe vocal effort mismatch between enrollment and test data causes a considerable decrease in recognition rates in automatic speaker recognition (Zhang and Hansen, 2007; Shriberg et al., 2008; Hanilçi et al., 2013b). To alleviate the problem, two main types of solutions have been proposed: (1) robust feature extraction methods and (2) techniques for compensating the standard features for the vocal effort mismatch. In the first group, Hanilçi et al. (2013a) compared several spectral estimation techniques to obtain Mel-frequency cepstral coefficients (MFCCs) more robust to mismatch in vocal effort. Results showed that *stabilized weighted linear prediction* (SWLP) performed better than the other candidates. Also *mixture* (Pohjalainen et al., 2014) and *power-law adjusted* linear prediction (LP) (Saeidi et al., 2016) have been shown to result in better acoustic features in mismatch conditions.

The second group of techniques are based in the domain of acoustic modeling and consist of different compensation algorithms to mitigate the undesirable effects of vocal effort mismatch. Motivated by the success of Gaussian mixture models (GMMs) in voice conversion (Stylianou et al., 1998), a GMM-based compensation of MFCCs was proposed by Hanilçi et al. (2013b). This technique improved the recognition rates in shouted versus normal mismatch conditions. Similar results were also reported by Ramírez López et al. (2017). A combination of both robust feature extraction and feature compensation is also possible. This kind of setup has been used, for instance, in the context of whispered speech where it improved speaker recognition accuracy in mismatched conditions (Fan and Hansen, 2009).

In this study, two compensation techniques are proposed for handling severe vocal effort mismatch (shouted versus normal) in a speaker identification task. The techniques are applied in the feature extraction stage of the speaker recognition system and their aim is to modify the spectral envelopes in shouts in such a way that the resulting acoustic features are better matched with the features extracted from normal speech in training. The techniques modify the power spectral estimate that is used in the MFCC computation with a shouted-to-normal compensation filter obtained using a GMM-based statistical mapping. One of the techniques aims to compensate for the changes in spectral tilt whereas the other technique uses a more general spectral model in the compensation. The proposed methods are evaluated in a shouted versus normal inset speaker identification task against three different reference techniques.

2. Shouted-to-normal vocal effort compensation

The vocal effort compensation is applied in the first stage of the MFCC (Davis and Mermelstein, 1980) chain, the computation of the signal's power spectrum. In the current study, the power spectrum of the MFCC chain is computed parametrically using LP as was done by Saeidi et al. (2016). Figs. 1 and 2 show the flow diagrams of the proposed two compensation methods: full envelope compensation (FEC) is shown in Fig. 1 and smoothed envelope compensation (SEC) is shown in Fig. 2. Both of the figures are divided into two parts: the training phase and the test phase. In the training phase, frames from parallel normal and shouted samples are aligned using dynamic time warping (DTW) and spectral features, denoted by $A^{\mathrm{N}}_{p_2}(z)$ and $A^{\mathrm{S}}_{p_2}(z)$ for normal and shouted speech, respectively, are computed. The spectral features parameterized are then used to train a joint-density GMM, used as a regression method for mapping shouted speech to normal speech. In the test phase, both methods take as an input a frame of shouted speech and yield as an output an all-pole spectral model, denoted by $B(z) = Z\{b(n)\}$, which is then used to compute the power spectrum input to the later stages of the MFCC chain. The aim of both methods is to compute such $B(z)$ that fits the spectral envelope of normal speech better than the spectral envelope computed from the original shouted speech frame.

Please cite this article as: E. Jokinen et al., Vocal effort compensation for MFCC feature extraction in a shouted versus normal speaker recognition task, Computer Speech & Language (2018), https://doi.org/10.1016/j.csl.2018.06.002