Digital Signal Processing 79 (2018) 142–151

Contents lists available at ScienceDirect

Digital Signal Processing

www.elsevier.com/locate/dsp

Studying the role of pitch-adaptive spectral estimation and speaking-rate normalization in automatic speech recognition

S. Shahnawazuddin [a,*], Nagaraj Adiga [b], Hemant K. Kathania [c], Gaydhar Pradhan [a], Rohit Sinha [d]

[a] Department of Electronics and Communication Engineering, National Institute of Technology, Patna, India
[b] Department of Computer Science, University of Crete, Greece
[c] Department of Electronics and Communication Engineering, National Institute of Technology, Sikkim, India
[d] Department of Electronics and Electrical Engineering, Indian Institute of Technology, Guwahati, India

ARTICLE INFO

Article history:
Available online xxxx

Keywords:
Pitch-adaptive spectral estimation
TANDEM STRAIGHT
Vocal-tract length normalization
Speaking-rate normalization
Glottal closure instants
Zero-frequency filtering

ABSTRACT

In the context of automatic speech recognition (ASR) systems, the front-end acoustic features should not be affected by signal periodicity (pitch period). Motivated by this fact, we have studied the role of pitch-synchronous spectrum estimation approach, referred to as TANDEM STRAIGHT, in this paper. TANDEM STRAIGHT results in a smoother spectrum devoid of pitch harmonics to a large extent. Consequently, the acoustic features derived using the smoothed spectra outperform the conventional Mel-frequency cepstral coefficients (MFCC). The experimental evaluations reported in this paper are performed on speech data from a wide range of speakers belonging to different age groups including children. The proposed features are found to be effective for all groups of speakers. To further improve the recognition of children's speech, the effect of vocal-tract length normalization (VTLN) is studied. The inclusion of VTLN further improves the recognition performance. We have also performed a detailed study on the effect of speaking-rate normalization (SRN) in the context of children's speech recognition. An SRN technique based on the anchoring of glottal closure instants estimated using zero-frequency filtering is explored in this regard. SRN is observed to be highly effective for child speakers belonging to different age groups. Finally, all the studied techniques are combined for effective mismatch reduction. In the case of children's speech test set, the use of proposed features results in a relative improvement of 21.6% over the MFCC features even after combining VTLN and SRN.

1. Introduction

Automatic speech recognition (ASR) is the task of generating textual output by decoding a digitally acquired acoustic pressure wave using machines (computers). With the introduction of hidden Markov models (HMM) for statistically learning the acoustic and linguistic attributes of speech [1], rapid progress has been made in ASR. In initial HMM-based systems, the observation densities for the HMM states were modeled using the continuous density Gaussian mixture models (GMM). Nowadays, the GMM is fast being replaced by deep neural networks (DNN) [2,3]. Since the computers available these days have excellent computing power, training very deep neural nets on large amounts of speech data for large vocabulary speech recognition task is no more a problem. Furthermore, fast and efficient techniques for training the network parameters have been developed [2]. As a result of the progress made in the research on speech processing, a number of speech-based user applications have been developed, e.g., voice-based web search, reading tutors, language learning tools and entertainment [4–6].

Another factor that plays an important role in ASR is the front-end speech parameterization module. The primary objective of front-end speech parameterization process is to extract the information relevant to the task and discard the rest. As a result, the front-end acoustic features result in a compact representation of raw speech signal. This significantly reduces the computational cost. The development of front-end speech parameterization techniques mimicking the human perception mechanism further aided in boosting the performance of ASR systems. Two dominant speech parameterization approaches used in ASR are the Mel-frequency cepstral coefficients (MFCC) [7] and perceptual linear prediction coefficients (PLPC) [8]. In last two decades, further advancements have taken place in the ASR domain which includes normalization

* Corresponding author.
E-mail addresses: s.syed@nitp.ac.in (S. Shahnawazuddin), nagaraj@csd.uoc.gr (N. Adiga), hemant.ece@nitsikkim.ac.in (H.K. Kathania), gdp@nitp.ac.in (G. Pradhan), rsinha@iitg.ernet.in (R. Sinha).

https://doi.org/10.1016/j.dsp.2018.05.003
1051-2004/

of the acoustic features prior to modeling by applying a set of linear transformations [9].

The performance of ASR systems deployed in the aforementioned user applications involving human machine interactions is affected by a number of factors. One among those is the inter-speaker variability such as age, gender, accent, speaking-rate, emotion and health conditions of the speakers contributing to the training and test speech. To impart robustness towards those factors, statistical models are trained on a large amount of speech data collected from a different class of speakers. In addition to that, techniques like feature-space maximum likelihood linear regression (fMLLR) [10] or vocal-tract length normalization (VTLN) [11] are commonly included to reduce the ill-effects of inter-speaker variability. Unfortunately, acoustic variations due to the aforementioned factors are so diverse that it is almost impossible to enhance the robustness towards all the factors simultaneously. In this paper, we attempt to reduce the ill-effects due to two of the dominant mismatch factors viz. the pitch differences among the speakers and the speaking-rate variability. In the following subsections, a brief review of existing research on the aforementioned aspects is presented.

1.1. Motivation and prior art

As mentioned earlier, MFCC is one of the most commonly employed front-end acoustic features. In the case of MFCC features, the speech signal is first analyzed into short-time overlapping frames. Next, the spectral representation is obtained by short-time Fourier transform (STFT). This is followed by warping the spectra to a non-uniform frequency scale by applying a triangular Mel-filterbank. Logarithmic compression of the filtered power spectrum is done next. The real cepstrum (RC) is then obtained by applying discrete cosine transform (DCT). The final MFCC features used for learning system parameters are obtained by low-time liftering of the cepstral coefficients.

In general, due to involved low-time liftering, it is expected that the MFCC features would be largely free from the effect of excitation. On the contrary, the MFCC features do get affected by the signal periodicity especially for the high-pitched speakers in comparison to the low-pitched ones [12,13]. The signal periodicity due to the excitation source is not well smoothed out while analyzing the signals having higher pitch values ($>$200 Hz) on warping of the frequency scale. This is mainly attributed to the narrow bandwidth ($\approx$100 Hz) of the lower-channel filters in the Mel-filterbank. Consequently, some ill-effects of the pitch (or signal periodicity) are still present in the derived features, especially for the high-pitched speakers, even after low-time liftering. Hence, the MFCC features exhibit enhanced variances for the higher-order coefficients for the high-pitched speakers in contrast to those for the low-pitched speakers. To highlight this phenomenon, a study conducted on voiced speech frames from different pitch groups for several different vowels was reported in [12]. We repeated that study for a single vowel[1] and the result for the same is summarized in Fig. 1. A mismatch in the variance of the higher-order coefficients (C_8–C_{12}) for the two pitch groups is easily noticeable. In [12], a more detailed study was reported quantifying the changes in the variance as well as the mean of the MFCC features on multiple vowels due to pitch. Furthermore, several other front-end cepstral features such as the linear prediction cepstral coefficient (LPCC), PLPC and the perceptual minimum variance distortionless response (PMVDR) were analyzed and found to be sensitive to the variation in the average pitch values [15,16].

[1] Since reliable vowel markings are available in the TIMIT database [14], this analysis was performed on the vowel data extracted from the same.

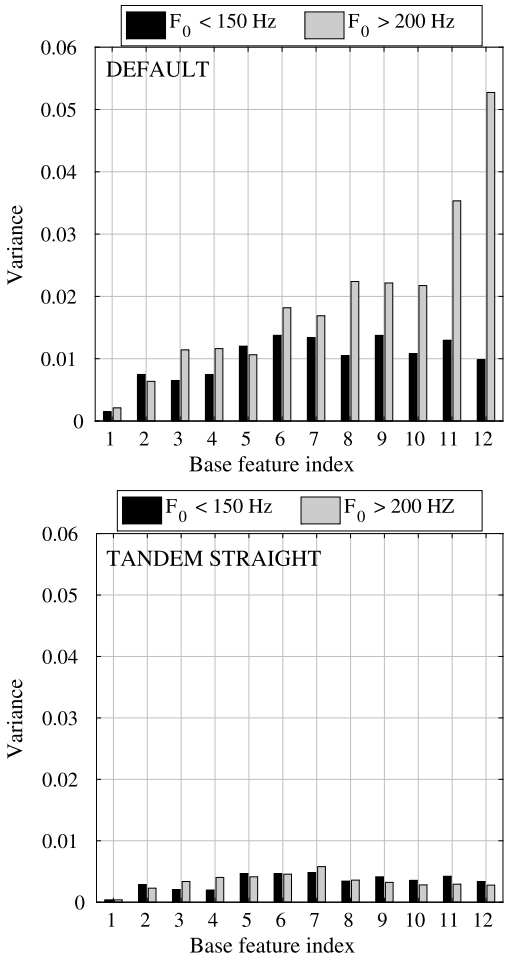

Fig. 1. Variance plots for the base MFCC features (C_1–C_{12}) for vowel /IY/ corresponding to two broad pitch (F_0) ranges. The feature vectors for nearly 2600 speech frames (for each group) corresponding to the central portion of the vowel were used for this analysis. For the chosen F_0 ranges, the mismatch in the variance of higher-order coefficients (especially C_{11} and C_{12}) is evident in the default case. The bottom pane highlights the reduction of variance mismatch with the TANDEM-STRAIGHT-based pitch-adaptive spectral estimation. These analyses were performed on the data extracted from TIMIT corpus. The feature vectors employed in these analyses have been normalized using cepstral mean and variance normalization.

The problem becomes much more challenging when we try to transcribe children's speech on ASR systems trained using adult's data and vice-versa. Despite the advances made in research on ASR, a severe degradation in recognition performance can still be noticed in such cases. The primary reason behind this observation is that the acoustic/linguistic properties of adults' and children's speech differ substantially due to morphological and physiological differences [17–20]. Consequently, achieving high recognition performance for both adult and child speakers on an ASR system becomes quite challenging. Several studies for addressing the acoustic mismatch in the context of children's ASR have been reported [21,22,13].

For high-pitched child speakers, the problem due to insufficient spectral smoothening becomes more pronounced. As argued in [12, 23], spectral smoothening can be increased by reducing the length of the low-time lifter. Even though such an approach improves the recognition performance for child speakers, there is a loss of relevant spectral information when a large number of cepstral coefficients are truncated. Motivated by that, low-rank feature projection was explored in [24]. In those works, the acoustic features were projected to lower-dimensional subspace prior to learning the model parameters. The low-rank projection matrix was derived ei-