Applied Acoustics 98 (2015) 52–61

Contents lists available at ScienceDirect

Applied Acoustics

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

A new pitch-range based feature set for a speaker's age and gender classification

Buket D. Barkana *, Jingcheng Zhou

Department of Electrical Engineering, University of Bridgeport, 221 University Ave., Bridgeport, CT, USA

ARTICLE INFO

Article history:
Received 13 May 2014
Received in revised form 20 April 2015
Accepted 21 April 2015
Available online 16 May 2015

Keywords:
Age and gender classification
Pitch range
Fundamental frequency
MFCCs

ABSTRACT

This paper presents a pitch-range (PR) based feature set for age and gender classification. The performance of the proposed feature set is compared with MFCCs, energy, relative spectral transform–perceptual linear prediction (RASTA_PLP), and fundamental frequency (F0). Voice activity detection (VAD) is performed to extract speech utterances before feature extraction. Two different classifiers, k-Nearest Neighbors (kNN) and Support Vector Machines (SVM) are used in order to evaluate the effectiveness of the feature sets. Experimental results are reported for the aGender database. Both kNN and SVM classifiers achieved the highest accuracy rates by the proposed PR feature set in age + gender and age classifications. PR features represent the pitch changes over time. In age + gender classification, the class of middle-aged female speaker is recognized with an accuracy of 92.86%, followed by senior female speakers with 83.61%, children with 83.02%, middle-aged male speakers with 73.58%, young female speakers with 67.35%, and senior male speakers with 34.33% by using 3PR features with the SVM classifier. Low classification accuracies are observed for young male speakers.

© 2015 Elsevier Ltd. All rights reserved.

1. Introduction

Detecting the age and gender of a speaker, given a short speech utterance, is a challenging task, and it is a rapidly emerging field of research because of the continually growing interest in applications of communication, human–computer interface and natural spoken-dialog systems. The understanding of the acoustics of an aging voice will provide a better understanding in a listener's perception of the aging voice. Human–computer interaction (HCI) systems such as a dialog system can be custom-designed, based on the speaker's voice portrait in order to guide the conversation and improve the level of customer satisfaction. The HMIHY system built by AT&T detects age and gender among various other voice signatures [1,2]. Gender detection is also a main part of the VIDIVIDEO European project, which deals with the semantic search of audio–visual documents [3].

There are many factors affecting the performance of such systems. The input speech material can be text-dependent or text-independent. In text-dependent systems, the speaker recites pre-determined words or phrases, whereas in text-independent systems there are no restrictions. Text-dependent systems give better performance than text-independent systems, because the speaker is aware of the task so that s/he is cooperative and consistent [4]. Another factor is the selection of the feature set to be used in the template representation. A desirable feature set should be easy to compute and robust. More importantly for text-independent systems, the feature set should also be unsusceptible to background noise since the speaker is often unaware of the task, and operating surroundings can have background noise and interference. The aim is to obtain a small and efficient set of acoustic features which represent the input pattern for the classification algorithms being trained [5].

This study focuses on the problem of extracting age and gender information from the speaker's voice. The speaker's gender is recognized in many automatic speech recognition (ASR) systems for the purpose of choosing gender-specific acoustic models. Gender classification has been the focus more than age classification in previous studies. Researchers have studied age identification on a small corpus [6,7]. Minematsu et al. stated that humans can identify age groups reliably even across different languages. One of these works in age detection performs a binary classification on a small corpus (elderly versus others) and achieves a 95% accuracy rate [8]. A method for detecting elderly speech based on prosodic features (jitter and shimmer) was proposed in [9].

Spectral and temporal feature sets, Mel-frequency cepstral coefficients (MFCCs), formant frequencies, fundamental frequency (F0),

* Corresponding author.
E-mail addresses: bbarkana@bridgeport.edu (B.D. Barkana), jinzhou@bridgeport.edu (J. Zhou).

http://dx.doi.org/10.1016/j.apacoust.2015.04.013
0003-682X/© 2015 Elsevier Ltd. All rights reserved.

energy, RASTA, jitter, shimmer, speech rate, harmony, and zero-crossing rates, were used to analyze speech characteristics in age and gender identification systems in previous studies. MFCCs offer a great deal of linguistic information for speech and speaker recognition applications [10]. Human perception of sound is based on a frequency analysis in the inner ear. MFCCs [11] are the cepstrum representations of this occurrence.

Zhan et al. compared the performance of Linear predictive coefficient (LPC), Linear-prediction cepstral coefficient (LPCC), MFCCs, and Bark-frequency cepstral coefficient (BFCC) feature sets by using a Gaussian mixture model (GMM), SVM, Multi-layer perception (MLP), kNN, and DS fusion classifiers in gender recognition. They reported that DS fusion classifiers achieved the highest gender classification accuracy of 93.07% female and 92.88% male, while the SVM classifier achieved 92.61% female and 92.28% male accuracy rates by using the MFCC feature set. It is also reported that LPC and LPCC sets achieved the lowest total classification accuracies for all classifiers [12].

Hu et al. proposed a two-level classifier with a pitch-based gender identification method to overcome the complexity of MFCC-based gender classification [13]. The first-stage classified the gender when pitch clearly indicates the gender of the speaker by using a threshold-based decision rule. The second-stage GMM classifier was used for undetermined speakers or difficult cases. They reported an accuracy rate of 98.65% for the TIDIGITS dataset.

Several speaker's age and gender classification studies are carried out by using aGender corpus. Ming et al. [38] proposed a method which combines five different acoustic level modeling methods as Gaussian Mixture Model (GMM) based on MFCC features, GMM–SVM mean supervector, GMM–SVM maximum likelihood linear regression (MLLR) supervector, GMM SVM Tandem supervector, and SVM baseline subsystems using 450-dimensional feature vectors including prosodic features. Their fusion system achieved 52.7% unweighted accuracy for the joint age and gender (age + gender) classification task and outperformed the GMM–MFCC system and SVM baseline, respectively, by 9.6% and 8.2%. Metze et al. [39] a comparative study in age and gender classification using aGender telephone speech corpus. They also compared the classification results with human performance on the same data. Four automatic classification methods, a parallel phone recognizer; dynamic Bayesian networks to combine prosodic features; linear prediction analysis; and GMM based on MFCC features are compared. Overall achieved accuracies were reported as 54%, 40%, 27%, and 42%, respectively. Overall classification accuracy by human listeners was reported as 55% for the aGender corpus. The classification of speakers' age and gender is a challenging task.

The calculation of MFCCs requires a large amount of storage space and has a high computation complexity. The performance of MFCC features is greatly affected by noisy recording environments. Although MFCCs are currently used for age and especially gender identification, temporal features of speech utterances may provide better information for age identification systems. We propose a PR feature set based on time–domain analysis. The proposed age and gender classification is based on three main steps: (1) pre-processing, which applies voice activity detection (VAD), (2) feature extraction, including MFCCs, Energy, RASTA_PLP, F0, and PR feature sets, and (3) classification. kNN and SVM classifiers are used in this stage to test the performance of selected combinations of the feature sets as well as individually is tested.

2. Database

In this work, data was taken from the aGender corpus [14,26]. It was supplied by the Interspeech 2010 Paralinguistic Challenge organization to support the development of speaker age and gender detection systems. The corpus consists of 49 h of telephone speech, stemming from 795 speakers, which are divided into a train (23 h, 471 speakers), development (14 h, 299 speakers) and test sets (12 h, 175 speakers) [14]. Four age groups make up the database: children, 7–14 years old (C); young-aged, 15–24 years old (YF/YM); middle-aged, 25–54 years old (MF/MM); and seniors, 55–80 years old (SF/SM). This choice was not motivated by physiological aspects that arise from the development of the human voice with increasing age, but rather it was based on market applications. Children are not subdivided as male and female speakers.

3. Methodology

Fig. 1 has three main steps. (1) The pre-processing step contains VAD to separate conversational speech from silence. There are many types of VAD algorithms. In this work, energy and zero-crossing rates [16] are used for VAD. (2) The feature extraction step calculates the feature sets used in classification. In addition to the calculation of well-known feature sets (MFCC, RASTA_PLP, F0), the proposed pitch-range (PR) feature set has been calculated. (3) The classification step uses kNN and SVM classifiers. kNN is a simple statistical learning algorithm in which an entity is classified by its neighbors. Computation time is short. SVM is a sophisticated supervised learning algorithm that requires training to determine the hyper-plane needed to separate classes accurately. The computation time of SVM can be long for multiple-class problems such as age and gender classification.

3.1. Pre-processing

The performance of speech processing applications is strongly affected by the quality of the speech signal. Although the speech signal is usually high-pass filtered to remove undesired low frequency components in practical speech applications, we do not do this in order to preserve spectral information that might be useful in age and gender classification. VAD is used in speech signal processing fields with the purpose of enhancing the quality of speech [15,16] before the feature extraction process. Short-time energy and zero-crossing measurements can be used in VAD. Fig. 2 shows short-time energy and zero-crossing rate of a speech utterance belonging to a child. A rectangular window of duration 32 ms (260 samples) is used with a 50% overlap.

An accurate and robust VAD plays an important role in the performance of the classifier. The theory of short-time energy and zero-crossing rate is briefly given below. The short-time energy is calculated as:

$$E = \sum_{n=0}^{N-1} |x(n)|^2 \tag{1}$$

where N is the window duration and $x(n)$ is the speech signal. Short-time zero-crossing rate is calculated as:

$$Z = \sum_{n=0}^{N-1} |sgn[x(n)] - sgn[x(n-1)]| \tag{2}$$

where $sgn[x(n)] = 1 \text{ if } x(n) \geqslant 0$ and $sgn[x(n)] = 1 \text{ if } x(n) < 0$.

3.2. Feature Extraction

Feature extraction is the process of calculating parameters that represent the characteristics of the input signal, whose output will have a direct and strong influence on the performance of classification systems. In this study, five different feature sets are calculated. They are MFCCs + energy (12 + 1 *features*), PLP-RASTA (13 *features*),

Download English Version:

https://daneshyari.com/en/article/754283

Download Persian Version:

https://daneshyari.com/article/754283

Daneshyari.com