Pattern Recognition Letters 84 (2016) 1–7

Contents lists available at ScienceDirect

Pattern Recognition Letters

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

# Feature optimisation for stress recognition in speech☆

Leandro D. Vignolo[a,*], S.R. Mahadeva Prasanna[b], Samarendra Dandapat[b], H. Leonardo Rufiner[a], Diego H. Milone[a]

[a] Research Institute for Signals, Systems and Computational Intelligence, sinc(i), FICH-UNL/CONICET, Argentina
[b] Department of Electronics and Communication Engineering, Indian Institute of Technology Guwahati, India

ARTICLE INFO

*Article history:*
Received 30 July 2015
Available online 27 July 2016

*Keywords:*
Evolutionary algorithms
Stressed speech
Emotional speech
Speech processing
Cepstral coefficients

ABSTRACT

Mel-frequency cepstral coefficients introduced biologically-inspired features into speech technology, becoming the most commonly used representation for speech, speaker and emotion recognition, and even for applications in music. While this representation is quite popular, it is ambitious to assume that it would provide the best results for every application, as it is not designed for each specific objective. This work proposes a methodology to learn a speech representation from data by optimising a filter bank, in order to improve results in the classification of stressed speech. Since population-based metaheuristics have proved successful in related applications, an evolutionary algorithm is designed to search for a filter bank that maximises the classification accuracy. For the codification, spline functions are used to shape the filter banks, which allows reducing the number of parameters to optimise. The filter banks obtained with the proposed methodology improve the results in stressed and emotional speech classification.



## 1. Introduction

The most widely used speech representation consists of the mel-frequency cepstral coefficients (MFCCs) [2,19], based on the linear voice production model, and uses a psycho-acoustic scale to mimic the frequency response in the human ear [9]. The MFCC features have been extensively used for speech [24,44], speaker [21], emotion [16,35,45] and language recognition [12], and even also for other applications not related to speech, such as music information retrieval [18]. However, the entire auditory system is not yet fully understood and the shape of the truly optimal filter bank is unknown. Moreover, the relevant part of the information contained in the signal depends on the application. Thus, it is unlikely that the same filter bank would provide the best performance for any kind of task. In fact, many alternative representations have been developed and some of them consist of modifications to the mel-scaled filter bank [44]. For example, a scheme for determining filter bandwidth was presented in [32], showing speech recognition improvements compared to traditional features. Also, auditory features based on Gammatone filters were developed for robust speech recognition [30]. Moreover, different approaches considering the noise energy on each mel band have been proposed in order to define MFCC weighting parameters [41,46]. The compression of filter bank energies according to the signal-to-noise ratio in each band was proposed in [15]. Similarly, other adjustments to the classical representation have been introduced [40]. Particularly for stressed speech classification, new time-frequency features have been presented [42]. Although these alternative features improve recognition results in particular tasks, to our knowledge, a methodology to automatically obtain an optimised filter bank for speech emotion classification has not been proposed.

Another common strategy that has been exploited for speech recognition is based on the optimisation of the feature extraction process in order to maximise the discrimination capability for a given corpus [7]. In this sense, the use of deep neural networks for learning filter banks was presented in [22], while other works introduced the use of linear discriminant analysis [6,43]. Genetic algorithms have also been applied for the design of wavelet-based representations [36]. Similarly, evolutionary strategies have been proposed for feature selection in other tasks [37]. Moreover, different approaches for the optimisation of speech features were based on evolutionary algorithms [38,39]. Also, an evolutionary approach for the generation of novel features has been proposed [25]. For stressed speech classification, genetic algorithms are also among the most successful feature selection techniques [8]. Nevertheless, there have not been attempts to optimise filter banks for the specific tasks of emotion or stress classification.

☆ This paper has been recommended for acceptance by J. Yang.

* Corresponding author at: Research Institute for Signals, Systems and Computational Intelligence, sinc(i), FICH-UNL/CONICET, Ciudad Universitaria CC 217, Ruta Nacional No 168 Km 472.4, (3000) Santa Fe, Argentina. Fax: +54 342 457 5228/5233.
*E-mail address:* ldvignolo@sinc.unl.edu.ar, ldvignolo@fich.unl.edu.ar (L.D. Vignolo).

http://dx.doi.org/10.1016/j.patrec.2016.07.017
0167-8655/

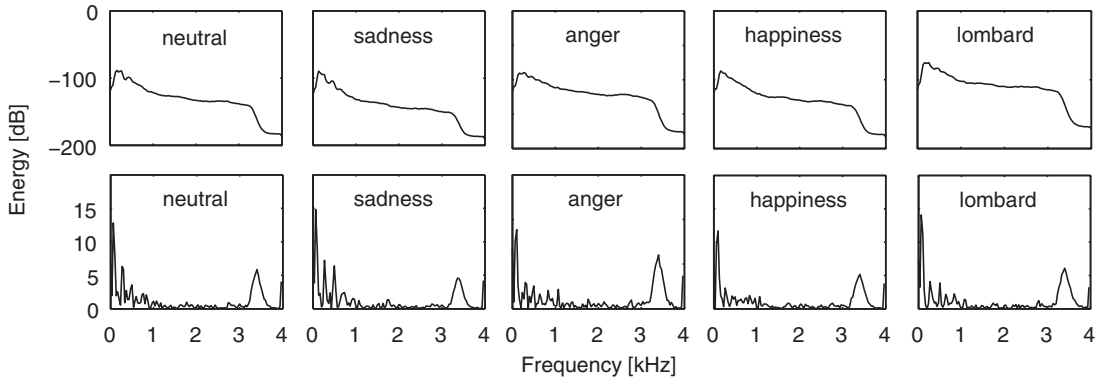


**Fig. 1.** Mean log spectrums (top) and first difference of mean log spectrums (bottom) for each of the five classes in a Hindi stressed speech corpus.

Evolutionary algorithms have proved to be effective in many complex optimisation problems [14]. Then, in order to tackle this challenging optimisation problem, we propose the use of an evolutionary algorithm for learning a filter bank from speech data. This work, based on the approach for the optimisation of filter banks, addresses the classification of different emotions and stress types in speech. The approach makes use of an evolutionary algorithm in order to optimise the filter bank involved in the extraction of cepstral features, with spline interpolation for parameter encoding. Our method attempts to provide an alternative speech representation to improve the classical MFCC on stress and emotion classification. A classifier is used to evaluate the evolved individuals, so that the accuracy is assigned as fitness. In contrast to previous work [39], in which the temporal dynamics of each class was modelled, for this task we introduced a static classification approach based on a single feature vector per utterance.

The remainder of this paper is organised as follows. In Section 2, a short overview of evolutionary algorithms is given, and also the feature extraction process for the MFCC is explained. Then, the proposal of this work is presented in Section 3 and the results obtained are discussed in Section 4. Finally, conclusions and proposals for future work are given in Section 5.

## 2. Background

### 2.1. Evolutionary algorithms

Evolutionary algorithms (EAs) are heuristic methods inspired by the process of biological evolution, which are useful for a wide range of optimisation problems [3,17,23]. The evolution is typically performed by means of natural operations like selection, mutation, crossover and replacement [4]. The selection operator assigns a reproduction probability to each individual in the population, favouring those with high fitness, in order to simulate natural selection. Mutation introduces random changes into chromosomes to maintain diversity within the population, while crossover combines information from parent individuals to create the offspring. Finally, the replacement strategy determines how many individuals in the current population are replaced by the offspring. This means that every population is replaced to improve fitness average and the loop is repeated to meet a stop criterion, after which the best individual provides an appropriate solution to the problem [10]. Solutions are represented by individuals and their information is coded by means of chromosomes, while their fitness is determined by a problem-specific objective function.

### 2.2. Mel-frequency cepstral coefficients

MFCCs are based on the linear speech production model, which assumes that the magnitude spectrum of a speech signal $S(f)$ can be formulated as the product of the excitation spectrum $X(f)$ and the frequency response of the vocal tract $H(f)$. That is $S(f) = X(f)H(f)$. Inspired on the human auditory system, the power spectrum is integrated into bands, according to the mel perceptual scale [9]. Given $M$ filters, $G_m(f)$, the energy outputs are computed by:

$$C(m) = \sum_f |S(f)|^2 G_m(f). \quad (1)$$

The logarithm is taken on the filter outputs, $C(m)$, and the MFCCs are computed by applying the discrete cosine transform (DCT) [9].

Even though these features are biologically inspired, their classification performance has been improved by other representations in different tasks. For example, a modification of MFCC that uses the known relationship between centre frequency and critical bandwidth was shown to increase noise robustness over traditional features in [32]. Also, [41] proposed performing Wiener filtering to mel sub-bands and estimating weights based on sub-band SNR-to-entropy ratio. Results showed that the method allows improving speech recognition performance in noisy environments. Furthermore, several experiments that compare the performance using different number of filters, filter shapes, filter spacing and spectrum warping were carried out [44]. In addition, the compression of filter bank energies according to the presence of noise in each mel sub-band was proposed so as to provide increased robustness in speech recognition and speaker identification [46].

In order to analyse the appropriateness of the mel filter bank for classification of stressed speech, we computed the mean of the log spectrum (MLS) along the frames (30 ms long) of the training utterances in each class. As it can be observed on top of Fig. 1, for a five-class corpus in Hindi language, the most discriminative information is found below 1 kHz, as the plots corresponding to different classes show different peaks within this band. Also, the first difference of each of the mean log spectrums was computed and is shown at the bottom of Fig. 1. These plots present peaks at high-frequency bands (from 3 to 4 kHz), showing different relative energy and shape, which could be useful for classification. This suggests that the mel filter bank is not entirely appropriate for this task. Fig. 2 shows the result of the same analysis performed on the FAU Aibo Emotion Corpus, which comprises recordings of German spontaneous speech [5,33]. As in the previous case, the most discriminative information seems to be found on lower frequency bands. However, for this corpus, the peaks are more prominent and the five emotions present noticeable different behaviour up to 2 kHz. Then, we can expect the optimum filter bank to be different for each corpus.

## 3. Evolutionary filter bank optimisation

Several parameters could be taken into account in the search for an optimal filter bank, such as the number of filters, filter shape