Feature optimisation for stress recognition in speech

Leandro D. Vignolo[1], S.R. Mahadeva Prasanna[2], Samarendra Dandapat[2], H. Leonardo Rufiner[1], and Diego H. Milone[1]

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

sinc(*i*) Research Center for Signals, Systems and Computational Intelligence (fich.unl.edu.ar/sinc)
L. D. Vignolo, S. R. M. Prasanna, S. Dandapat, H. L. Rufiner & D. H. Milone; "Feature optimisation for stress recognition in speech"
Pattern Recognition Letters, Vol. 84, pp. 1-7, dec, 2016.

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) [Albornoz et al., 2011, Reyes-Vargas et al., 2013], based on the linear voice production model, and uses a psycho-acoustic scale to mimic the frequency response in the human ear [Deller et al., 1993]. The MFCC features have been extensively used for speech [Sarikaya and Hansen, 2000, Zheng et al., 2001], speaker [Sahidullah and Saha, 2013], emotion [Ooi et al., 2014, Ververidis and Kotropoulos, 2006, Zheng et al., 2014] and language recognition [Huang et al., 2013], and even also for other applications not related to speech, such as music information retrieval [Qin et al., 2013]. 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 [Zheng et al., 2001]. For example, a scheme for determining filter bandwidth was presented in [Skowronski and Harris, 2004], showing speech recognition improvements compared to traditional features. Also, auditory features based on Gammatone filters were developed for robust speech recognition [Shao et al., 2009]. Moreover, different approaches considering the noise energy on each mel band have been proposed in order to define MFCC weighting parameters [Yeganeh et al., 2008, Zhou et al., 2007]. The compression of filter bank energies according to the signal-to-noise ratio in each band was proposed in [Nasersharif and Akbari, 2007]. Similarly, other adjustments to the classical representation have been introduced [Wu and Cao, 2005]. Particularly for stressed speech classification, new time-frequency features have been presented [Zão et al., 2014]. 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 [Bŏril, H. and Fousek, P. and Pollák, P., 2006]. In this sense, the use of deep neural networks for learning filter banks was presented in [Sainath et al., 2013], while other works introduced the use of linear discriminant analysis [Burget and Heřmanský, 2001, Zamani et al., 2011]. Genetic algorithms have also been applied for the design of wavelet-based representations [Vignolo et al., 2013a]. Similarly, evolutionary strategies have been proposed for feature selection in other tasks [Vignolo et al., 2013b]. Moreover, different approaches for the optimisation of speech features were based on evolutionary algorithms [Vignolo et al., 2011a, Vignolo et al., 2011b]. Also, an evolutionary approach for the generation of novel features has been proposed [Schuller et al., 2006]. For stressed speech classification, genetic algorithms are also among the most successful feature selection techniques [Casale et al., 2007]. Nevertheless, there have not been attempts to optimise filter banks for the specific tasks of emotion or stress classification.

Evolutionary algorithms have proved to be effective in many complex optimisation problems [Lin et al., 2015]. 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 [Vignolo et al., 2011b], 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 [Andrews and McNicholas, 2013,Paul and Das, 2015, Sanchez-Diaz et al., 2014]. The evolution is typically performed by means of natural operations like selection, mutation, crossover and replacement [Bäck et al., 1997]. 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 [Engelbrecht, 2007]. 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

sinc(*i*) Research Center for Signals, Systems and Computational Intelligence (fich.unl.edu.ar/sinc)
L. D. Vignolo, S. R. M. Prasanna, S. Dandapat, H. L. Rufiner & D. H. Milone; "Feature optimisation for stress recognition in speech"
Pattern Recognition Letters, Vol. 84, pp. 1-7, dec, 2016.

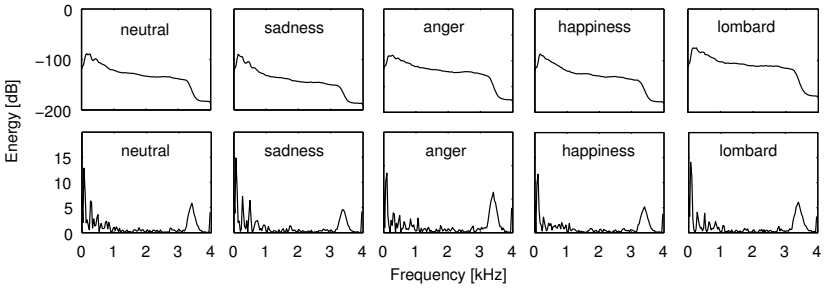

Figure 1: Mean log spectrums (top) and first difference of mean log spectrums (bottom) for each of the five classes in a Hindi stressed speech corpus.

scale [Deller et al., 1993]. 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) [Deller et al., 1993].

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 [Skowronski and Harris, 2004]. Also, [Yeganeh et al., 2008] 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 [Zheng et al., 2001]. 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 [Zhou et al., 2007].

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 Figure 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 Figure 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. Figure 2 shows the result of the same analysis performed on the FAU Aibo Emotion Corpus, which comprises recordings of German spontaneous speech [Batliner et al., 2008,Steidl, 2009]. 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 and filter gain. However, as the number of parameters is increased, the problem becomes extremely complex, so there is a tradeoff between optimisation complexity and flexibility. In previous works, three parameters were considered for each triangular filter in the filter bank; these correspond to the frequency values where: the triangle begins, reaches its maximum and ends. The results suggested that this formula-

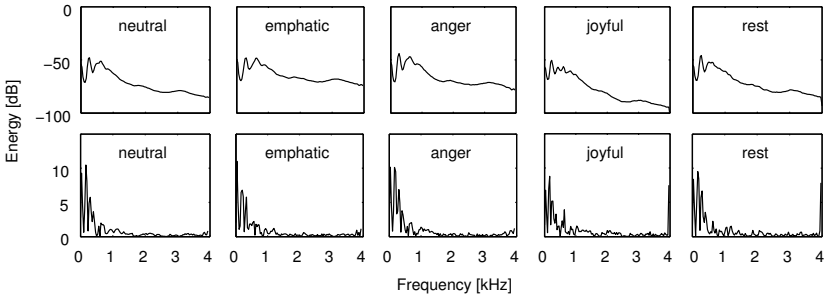

Figure 2: Mean log spectrums (top) and first difference of mean log spectrums (bottom) for each of the five classes in the FAU Aibo emotion corpus.

tion produced an ill-conditioned problem [Vignolo et al., 2011a]. This approach, referred to as *Evolutionary Spline Cepstral Coefficients* (ESCCs), uses splines to shape the filter banks. In this way, the chromosomes hold the spline parameters instead of the filter bank parameters, so that the chromosome size and the search space are reduced. Splines are chosen because they allow limiting the start and end points of the functions' domain easily, which is useful to generate filter banks that cover the frequency range of interest.

We defined a spline mapping $y = c(x)$, with $y \in [0, 1]$, and x taking n_f equidistant values in $(0, 1)$. Then, for a filter bank with n_f filters, x_i was assigned to filter i, with $i = 1, ..., n_f$. For a given chromosome, all y_i values corresponding to x_i were obtained by cubic spline interpolation. We used two different splines. The first one was used to determine the frequency values corresponding to the maximum of each triangular filter, from 0 Hz to half the sampling frequency (f_s). The edge points of each filter were fixed to the frequencies where its adjacent filters were maximum, thereby determining filter overlapping as well. The second spline was used to set the amplitude of each filter.

3.1 Optimisation of filter frequency locations

Here we used a monotonically increasing spline, constrained to $c(0) = 0$ and $c(1) = 1$. We set four parameters to define the spline I: y_1^I and y_2^I corresponding to fixed values x_1^I and x_2^I, and the derivatives, σ and ρ, at the fixed points $(x = 0, y = 0)$ and $(x = 1, y = 1)$. It is important to point out that parameter y_2^I was restricted to be equal to or greater than y_1^I, in order to obtain monotonically increasing splines. Then, parameter y_2^I was obtained as $y_2^I = y_1^I + \delta_{y_2}$, and the parameters actually coded in the chromosomes were y_1^I, δ_{y_2}, σ and ρ. Given a particular chromosome, which set the values for these parameters, the $y[i]$ corresponding to the $x[i]$ $\forall\, i = 1, ..., n_f$ were obtained by spline interpolation. This is schematised in Figure 3.

The $y[i]$ values obtained by spline interpolation were linearly mapped to the frequency range of interest (from 0 Hz to $f_s/2$), so the frequency values for the maximum of each of the n_f filters, f_i^c, were obtained as

$$f_i^c = \frac{(y[i] - y_m) f_s}{y_M - y_m}, \tag{2}$$

where y_m and y_M are the spline minimum and maximum values, respectively. In this way, in the segments where the slope of the spline was low, the filters were far from each other; and where the slope was high, the filters were closer. There was also a parameter $0 < a < 1$ to limit the range of y_1^I and y_2^I to $[a, 1-a]$, with the purpose of keeping the splines within $[0, 1]$. The splines that went beyond these boundaries were modified, fixing them to 0 or 1.

3.2 Optimisation of filter amplitudes

The amplitude spline had the only restriction of lying in the range $[0, 1]$, but y was free at $x = 0$ and $x = 1$. Therefore, the parameters to be optimised here were the y values y_1^{II}, y_2^{II}, y_3^{II} and y_4^{II}, corresponding to the fixed x values x_1^{II}, x_2^{II}, x_3^{II} and x_4^{II}. These four y_j^{II} were limited to $[0, 1]$. In this way, we obtained n_f interpolated values that were used to set the gain of the filters. This is shown in Figure

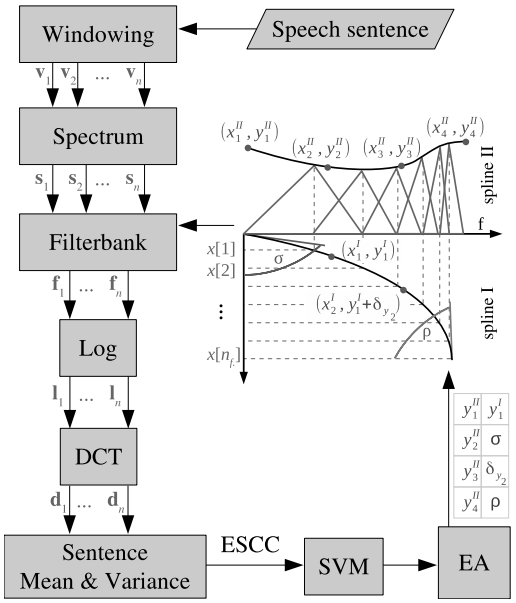

Figure 3: General scheme explaining the optimisation strategy and the feature extraction process. The output vectors of each block, $\mathbf{s}_i$, $\mathbf{f}_i$, $\mathbf{l}_i$ and $\mathbf{d}_i$, indicate that each window $\mathbf{v}_i$ is processed isolated and, finally, the mean and variance for each coefficient is computed from the $\mathbf{d}_i$ vectors.

3, where the gain of each filter was set according to spline II. Hence, the assumption is that evolutionary optimisation will enhance the frequency bands that are relevant for the classification of stressed and emotional speech.

3.3 Overview of the optimisation process

Every individual contains a set of spline parameters, encoding a particular filter bank. The selection process takes into account the classification performance. Prior to the classification, the filter bank proposed by a chromosome was used for processing the signals in the data set, in order to obtain the speech features (Figure 3). Then, the fitness assigned to an individual was the accuracy rate obtained. To avoid overfitting of the filter banks, the training and validation partitions used during the optimisation were randomly reassembled on every generation. As this strategy can slow down convergence, we introduced an alternative for the fitness computation that counterbalances this side effect. This consists in computing the fitness of an individual surviving for more than one generation as an average of the performances on the generations it remained unchanged. If the chromosome was altered, the performance obtained in the current generation is assigned.

For the joint optimisation of filter amplitudes and positions, the proposed codification makes possible to reduce the chromosome size from $2n_f$ to the number of spline parameters. Here we used 26 filters, so the size of the chromosome was reduced from 46 to 8. It is important to note that, using the spline codification, the length of the chromosome does not depend on the size of the filter bank. Chromosomes were then coded as strings of 8 real numbers (4 parameters for each of the two splines), which were randomly initialized using uniform distribution.

In this EA, tournament selection and standard one-point crossover methods were used, while the mutation operator was designed to modify splines parameters. The parameters were randomly chosen by the operator and the modifications were performed using a uniform random distribution. In order to favour convergence, the elitist strategy was incorporated into the search, which means that the best individual from one generation was maintained into the next one [Engelbrecht, 2007]. To determine the proper values for the parameters of the evolutionary algorithm, preliminary optimisation experiments were performed using a separate subset of the Hindi corpus. Based on the analysis of the evolution of the population in these experiments, a proper set of parameters was set as follows. The size of the population was set to 30 individuals, while crossover and mutation probabilities were set to 0.9 and 0.12, respectively. A maximum of 300 generations was set as stopping criteria; however, the evolution was early stopped when there was no improvement during a lapse of 10% of the maximum number of generations.

4 Results and discussion

4.1 Materials

In the experiments, the FAU Aibo Emotion Corpus [Batliner et al., 2008, Steidl, 2009] and a simulated stressed speech corpus in Hindi language

sinc(*i*) Research Center for Signals, Systems and Computational Intelligence (fich.unl.edu.ar/sinc)
L. D. Vignolo, S. R. M. Prasanna, S. Dandapat, H. L. Rufiner & D. H. Milone; "Feature optimisation for stress recognition in speech"
Pattern Recognition Letters, Vol. 84, pp. 1-7, dec, 2016.

[Shukla et al., 2011] were used. The Hindi language database consists of stressed speech signals recorded from fifteen speakers, ten male and five female. The speech utterances were sampled at 8 kHz and include *neutral* speech and four *acted* stress conditions: *anger*, *happiness*, *lombard* and *sadness*. Each recorded signal consisted of a keyword, which was uttered within a sentence and then isolated. For each of these five classes, 395 instances were used during the optimisation (80% for training and the remainder for validation), and other 99 examples were left for the final test (i.e. a total of 495 utterances for testing).

The FAU Aibo corpus includes speech in *neutral* state and four classes of spontaneous emotions: *anger* (angry, touchy and reprimanding), *emphatic*, *positive* (motherese and joyful) and *rest*. It provides clearly-defined test and training partitions with speaker independence and different room acoustics. The speech utterances in this corpus, which consist of sentences of varying lengths, were recorded at 16 kHz sampling rate. In order to compare filter banks optimised for both corpora, we decided to sub-sample the recordings to 8 kHz. The speaker-independent training and test partition suggested for the InterSpeech 2009 Emotion Challenge [Schuller et al., 2009] was used. Therefore, a set consisting of 9959 instances was used for the optimisation (80% for training and 20% for validation) and 8257 instances were used for the final test.

4.2 Feature extraction and experimental setup

For each candidate filter bank, speech signals were processed on a frame basis, using a 30-ms Hamming window with 7.5-ms step. This is a common setup for speech processing [Deller et al., 1993] and provided good results in preliminary tests with stressed speech. All frames without speech activity were discarded and the different filter banks were applied to voiced frames, in order to compute 13 cepstral coefficients from the filter output energies. As opposed to our previous work [Vignolo et al., 2011b], in which we used a temporal modelling approach for phoneme classification, here the feature vectors extracted from the frames of an utterance were averaged. This means that for each sentence we obtained a single feature vector, which allowed using a static classifier as shown in Figure 3. In addition, the mean, minimum and maximum frame energy were appended, obtaining a vector of 16 features. Besides, for the experiments with the FAU Aibo corpus, the variances of the cepstral coefficients were also appended, which allowed to obtain satisfactory baseline results in this case. No tuning was performed on the hyper-parameters for this corpus, that is, the same values used for Hindi stressed speech were set.

The classifier was trained using 80% of the instances in the training set, and it was then evaluated on the remaining 20%. This provided a good tradeoff between the number of training instances and the number of test instances required to evaluate the statistical significance of results. This test set was separated and it was only used to evaluate the best solution after the optimisation. As mentioned in Section 3.3, the data partition used during the optimisation was reassembled randomly on every generation so as to avoid overfitting. The performance of the best filter bank provided by the EA was evaluated using this separate test set and a classifier based on Support Vector Machines (SVMs) with polynomial kernel [Steinwart and Christmann, 2008]. As the FAU Aibo corpus has unbalanced classes, we used the *Unweighted Average Recall* (UAR) [Rosenberg, 2012] in order to measure and compare the classification performance with the test set. Also, the training set was resampled with replacement in order to obtain the same number of instances for each class.

4.3 Filter bank optimisation

Table 1 presents the classification results obtained on the test sets using SVMs, comparing the performance of optimised features and MFCCs. ESCC-Eh and ESCC-Ef correspond to the filter banks optimised for Hindi stressed speech and FAU Aibo Emotion Corpus, respectively. Although the performances obtained with MFCC are acceptable for both corpora, ESCC-Eh and ESCC-Ef provide significant improvements in UAR for their corresponding tasks. Moreover, ESCC-Eh and ESCC-Ef outperformed the standard features in the recognition of most of the stress and emotion classes. The statistical significance of the results obtained was evaluated by estimating the probability for the optimised features to perform better than the MFCC.

Table 1: Results obtained on the final classification tests [%].

Hindi corpus	MFCC	ESCC-Eh	FAU Aibo	MFCC	ESCC-Ef
neutral	78.79	**95.96**	neutral	31.82	**39.22**
anger	82.83	**89.90**	anger	46.81	**52.86**
happiness	**87.88**	84.85	emphatic	**39.32**	39.19
lombard	74.75	**89.90**	positive	57.67	**65.12**
sadness	83.84	**95.96**	rest	**24.54**	16.12
UAR	81.62	**91.31**	UAR	40.03	**42.50**

To estimate this probability, the error distributions for the optimised features (p_1) and MFCC (p_2) were approximated. Although these follow a binomial distribution, for a large number of trials they can be approximated with Gaussian distributions. Then, the probability for the optimised features to perform better than MFCC, $Pr(p_1 < p_2)$, was estimated by integrating the distribution $p(p_1, p_2)$ in the region $p_1 < p_2$. The probabilities that ESCC-Eh and ESCC-Ef perform better than MFCC on their corresponding tasks were 0.9999 and 0.9996, respectively.

Another point to remark is that only 183 and 90 generations were needed in order to obtain ESCC-Eh and ESCC-Ef, respectively. The time required for the algorithm to converge was 22 hs to obtain ESCC-Ef, and 44 hs and 30 minutes for ESCC-Eh, using a personal computer with an Intel Core I7 processor and 8 GB RAM.

We performed the same tests, in stress and emotion classification, using a filter bank optimised for phoneme recognition in [Vignolo et al., 2011b] (ESCC-Ph in Figure 5). Moreover, we evaluated the performance of ESCC-Eh with the German emotion corpus and, conversely, the performance of ESCC-Ef on the Hindi stressed speech corpus. These results are shown in Table 2. As expected, ESCC-Ph is not useful for stressed speech classification. Furthermore, the results obtained with ESCC-Eh and ESCC-Ef on both corpora suggest that, even with different languages and classes, both tasks are closely related.

Table 3 shows more detailed information about the classification performance comparing MFCC and the optimised filter banks. In these confusion matrices, rows correspond to the actual class and columns correspond to the predicted class, while the percentages of correct classification lay on the diagonal. These matrices show coincidence between the classes that are more confused using MFCCs, and the ones that are more confused using ESCC-Eh and ESCC-Ef, respectively. However, it can be noticed that the optimised features allowed reducing most of the percentages outside the diagonal. For example, for the FAU Aibo corpus, all the off-diagonal values in columns `e` and `r` were significantly reduced with ESCC-Ef. It can be also noticed that the performance for *happiness* was not improved, and an explanation could be obtained from Figure 1 as follows. The first difference of the MLS for this class presents lower and fewer peaks at low frequencies (0-500 Hz), which seems to be the band enhanced by ESCC-Eh (Figure 5). This suggests that the evolution of the filter bank could have faced a trade-off, in which a better average result was obtained by enhancing this band, despite of a lower accuracy for *happiness*. A similar case arises for class *rest* in the FAU Aibo corpus. This class includes different types of emotions and contains few examples, making it difficult to model. Thus, it could be expected that a fine tuning for improving the overall UAR could deteriorate the performance for this class.

We performed another experiment in order to evaluate speaker independence with the Hindi corpus, which contains speech from 15 speakers. Then we assessed the performance of the SVM classifier through cross validation with 15 different data partitions. That is, on each partition we separated all the sentences corresponding to one speaker for testing and the rest for training. However, since ESCC-Eh was optimised using training utterances from all speakers, this step in the validation methodology would not be speaker-independent. Therefore, in this test we used the ESCC-Ef, which was optimised using another corpus. The unweighted accuracy obtained with the optimised representation was higher than the result obtained with MFCC (42.38% and 38.05%, respectively). Moreover, since ESCC-Ef was optimised for a different corpus, the

sinc(*i*) Research Center for Signals, Systems and Computational Intelligence (fich.unl.edu.ar/sinc)
L. D. Vignolo, S. R. M. Prasanna, S. Dandapat, H. L. Rufiner & D. H. Milone; "Feature optimisation for stress recognition in speech"
Pattern Recognition Letters, Vol. 84, pp. 1-7, dec, 2016.

Table 2: Results obtained on the final classification tests [UAR%].

	MFCC	ESCC-Ph	ESCC-Eh	ESCC-Ef
Hindi corpus	81.62	42.22	**91.31**	78.59
FAU Aibo	40.03	31.51	38.47	**42.50**

evaluation in the speaker-independent condition revealed that the proposed methodology is able to provide more general features, which are also useful in other corpora for similar tasks on emotion recognition. Although the performances obtained in the speaker-dependent condition (Table 2) suggest that MFCC could provide a more general solution for different corpora, the results discussed above show that it is also possible to obtain generalisation through different corpora with this methodology. Nevertheless, the motivation of this work is to provide a methodology to fit a filter bank for a particular task, which allows improving classification results in the application of interest.

In order to provide a tool for a qualitative analysis of these results, a Self-Organising Map (SOM) [Kohonen, 2000] was trained to show the topological distribution of the stress classes from the Hindi speech corpus in two dimensions. SOMs are trained in an unsupervised manner and preserve the neighbourhood relations of the input space. For each training case, the ESCC-Eh feature vectors were used as input of a 25-neuron network, and the same process was repeated for MFCC. Figure 4 presents the SOM obtained, showing that the ESCC-Eh allows grouping stress classes *happiness* (H) and *lombard* (L) into one cluster each. Conversely, the map corresponding to MFCC presents a more complex topology, in which more classes are split into separate clusters. The exceptions are classes *neutral* (N) and *sadness* (S), which are grouped in fewer clusters in the map corresponding to MFCC. However, Table 3 shows that these two classes are better discriminated in the new feature space, and they exhibit the highest improvements of ESCC-Eh with respect to MFCC. This suggests that the features provided by the optimised filter bank make up a representation in which the stress classes can be more easily discriminated.

The performance of the optimised representations was also compared with other state-of-the-art features, including: the InterSpeech 2009 Emotion Challenge feature set consisting of 384 attributes (IS09 Emotion) [Schuller et al., 2009], the emobase2010 feature subset based on the Interspeech 2010 Paralinguistic Challenge (IS10 Emobase, 1582 attributes) [Schuller et al., 2010], the feature set proposed for the InterSpeech 2011 Speaker State Challenge (IS11 Speaker State, 4369 attributes) [Schuller et al., 2011] and the feature of the InterSpeech 2013 Computational Paralinguistics Challenge (IS13 ComParE, 6374 attributes) [Schuller et al., 2013]. All of these feature sets have been extracted with the openSMILE library [Eyben et al., 2010]. Also, features based on the MLS proposed by [Albornoz et al., 2011] were included in the comparison, using the first 30 (MLS30) coefficients (covering from 0 to 1200 Hz) and these MLSs with the 30 standard deviation coefficients (MLS30+30) [Albornoz and Milone, 2015]. Results are presented in Table 4. It is interesting to note that the performance ranking of the feature sets from InterSpeech challenges differs between Hindi and FAU Aibo databases. The MLS performed well in both databases when the variances were included. For the Hindi corpus, the IS13 ComParE features performed better than MFCCs. However, the ESCC-Eh and ESCC-Ef provided the best performances for the Hindi and FAU Aibo databases, respectively. The statistical significance was evaluated as previously described; for the Hindi corpus the probability that ESCC-Eh performs better than the features of IS13 ComParE was 0.9996, and for the FAU Aibo corpus the probability that ESCC-Ef performs better than MFCC was also 0.9996.

Figure 5 shows the standard mel filter bank (MFCC), together with the filter banks optimised for phoneme classification, Hindi stressed speech and emotional German speech, respectively. It can be noticed that the optimised filter banks differ widely from mel filter bank. On the one hand, ESCC-Ph provides higher resolution in the appropriate frequency band, assigning almost equal weight to all filters. On the other hand, ESCC-Eh and ESCC-Ef provide higher resolution and gain on narrow bands. Moreover, the ESCC-Eh enhances

Table 3: Results obtained on the final classification tests [UAR%].

		MFCC					ESCC-Eh				
		a	h	l	n	s	a	h	l	n	s
Hindi corpus	a= anger	82.83	10.10	07.07	00.00	00.00	89.90	05.05	04.04	01.01	00.00
	h= happiness	01.01	87.88	03.03	07.07	01.01	02.02	84.85	02.02	06.06	05.05
	l= lombard	09.09	09.09	74.75	07.07	00.00	06.06	01.01	89.90	03.03	00.00
	n= neutral	00.00	13.13	05.05	78.79	03.03	00.00	02.02	01.01	95.96	01.01
	s= sadness	00.00	11.11	00.00	05.05	83.84	00.00	03.03	00.00	01.01	95.96
		MFCC					ESCC-Ef				
		a	e	n	p	r	a	e	n	p	r
FAU Aibo	a= anger	46.81	19.15	10.47	02.62	20.95	52.86	13.09	14.57	05.73	13.75
	e= emphatic	14.26	39.32	21.35	05.77	19.30	16.45	39.19	30.77	04.91	08.69
	n= neutral	10.32	14.90	31.82	19.90	23.06	12.29	12.68	39.22	22.82	12.98
	p= positive	03.26	05.58	11.16	57.67	22.33	03.26	03.72	10.23	65.12	17.67
	r= rest	10.07	06.41	23.44	35.53	24.54	13.37	06.78	29.85	33.88	16.12

Table 4: Performance comparison with state-of-the-art features [UAR%].

Hindi		FAU Aibo	
IS09 Emotion	70.71	IS09 Emotion	37.29
IS10 Emobase	79.80	IS10 Emobase	35.43
IS11 Speaker State	82.83	IS11 Speaker State	32.64
IS13 ComParE	84.04	IS13 ComParE	32.48
MLS30	48.69	MLS30	33.59
MLS30+30	61.21	MLS30+30	36.48
MFCC	81.62	MFCC	40.03
ESCC-Eh	**91.31**	ESCC-Ef	**42.50**

both the lowest and the highest frequency bands, which agree with the observations made based on Figure 1. ESCC-Ef also emphasises low frequency bands, although the amplitude of its filters decreases beyond 2000 Hz and it does not suppress mid-band frequencies as much as ESCC-Eh. This is consistent with our previous analysis based on Figures 1 and 2. Furthermore, these results suggest that our methodology is able to provide filter banks that capture relevant information from particular frequency bands.

5 Conclusion and future work

In this work, an evolutionary optimisation method has been proposed in order to improve stressed and emotional speech classification results. The chromosome codification based on splines allowed reducing the number of optimisation parameters, while maintaining the quality and diversity of possible solutions. This encoding also helped to simplify the filter bank optimisation problem, making possible to speed up the convergence of the EA to decent solutions. Also, we proposed a static classification scheme based on mean and variance features to describe utterances with a single vector, which allowed simplifying the optimisation problem while providing satisfactory performances.

The performances obtained in two different speech corpora show that the approach is useful to find an improved speech representation for the classification of stressed speech. These results also suggest that there is further room for improvement over the classical filter bank on specific tasks.

It is important to note that our study was limited to clean speech signals; however, the impact of noise on the shape of the filter banks should be evaluated. Thus, further experiments will include noisy signals, as well as other types of stressed speech and different languages. In addition, there are other filter bank parameters, such as the filter bandwidth, that were not optimised in this work and will be taken into account in the future. In order to obtain improvements on the individual accuracies for all classes, a multi-objective evolutionary algorithm could be used to consider the accuracy for each class as a separate objective.

Acknowledgements

This work was supported by the Argentinian Ministerio de Ciencia, Tecnología e Innovación Productiva and by the Indian Department of Science and Technology, under project IN1103. Also, the authors wish to acknowledge the support provided by Agencia Nacional de Promoción Científica y Tecnológica (with projects PICT 2011-2440, PICT

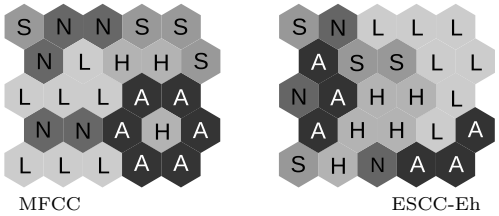

Figure 4: SOM clustering obtained on Hindi corpus for MFCC and ESCC-Eh, respectively.

2014-1442), Universidad Nacional de Litoral (with projects CAID 2011-519, -525 and PACT 2011-058) and Consejo Nacional de Investigaciones Científicas y Técnicas from Argentina.

References

[Albornoz and Milone, 2015] Albornoz, E. and Milone, D. (2015). Emotion recognition in never-seen languages using a novel ensemble method with emotion profiles. *IEEE Transactions on Affective Computing,*, PP(99):1–1.

[Albornoz et al., 2011] Albornoz, E. M., Milone, D. H., and Rufiner, H. L. (2011). Spoken emotion recognition using hierarchical classifiers. *Computer Speech and Language*, 25(3):556–570.

[Andrews and McNicholas, 2013] Andrews, J. L. and McNicholas, P. D. (2013). Using evolutionary algorithms for model-based clustering. *Pattern Recognition Letters*, 34(9):987–992.

[Bäck et al., 1997] Bäck, T., Hammel, U., and Schewfel, H.-F. (1997). Evolutionary computation: Comments on history and current state. *IEEE Trans. on Evolutionary Computation*, 1(1):3–17.

[Batliner et al., 2008] Batliner, A., Steidl, S., Hacker, C., and Nöth, E. (2008). Private emotions versus social interaction: a data-driven approach towards analysing emotion in speech. *User Modeling and User-Adapted Interaction*, 18(1-2):175–206.

[Burget and Heřmanský, 2001] Burget, L. and Heřmanský, H. (2001). Data Driven Design of Filter Bank for Speech Recognition. In *Text, Speech and Dialogue*, Lecture Notes in Computer Science, pages 299–304. Springer.

[Bŏril, H. and Fousek, P. and Pollák, P., 2006] Bŏril, H. and Fousek, P. and Pollák, P. (2006). Data-Driven Design of Front-End Filter Bank for Lombard Speech Recognition. In *Proc. of INTERSPEECH 2006 - ICSLP*, pages 381–384, Pittsburgh, Pennsylvania.

[Casale et al., 2007] Casale, S., Russo, A., and Serrano, S. (2007). Multistyle classification of speech under stress using feature subset selection based on genetic algorithms. *Speech Communication*, 49(10-11):801–810.

[Deller et al., 1993] Deller, J. R., Proakis, J. G., and Hansen, J. H. (1993). *Discrete-Time Processing of Speech Signals*. Macmillan Publishing, NewYork.

[Engelbrecht, 2007] Engelbrecht, A. (2007). *Computational Intelligence: An Introduction*. Wiley.

[Eyben et al., 2010] Eyben, F., Wöllmer, M., and Schuller, B. (2010). Opensmile: The Munich versatile and fast open-source audio feature extractor. In *Proc. of the Int. Conf. on Multimedia*, MM '10, pages 1459–1462, New York, NY, USA. ACM.

[Huang et al., 2013] Huang, C.-L., Matsuda, S., and Hori, C. (2013). Feature normalization using MVAW processing for spoken language recognition. In *Signal and Information Processing Association Annual Summit and Conference (APSIPA), 2013 Asia-Pacific*, pages 1–4.

[Kohonen, 2000] Kohonen, T. (2000). *Self-Organizing Maps*. Springer series in information sciences, 30. Springer, 3rd edition.

[Lin et al., 2015] Lin, C. D., Anderson-Cook, C. M., Hamada, M. S., Moore, L. M., and Sitter, R. R. (2015). Using genetic algorithms to design experiments: A review. *Quality and Reliability Engineering International*, 31(2):155–167.

[Nasersharif and Akbari, 2007] Nasersharif, B. and Akbari, A. (2007). SNR-dependent compression of enhanced Mel sub-band energies for compensation of noise effects on MFCC features. *Pattern Recognition Letters*, 28(11):1320–1326. Advances on Pattern recognition for speech and audio processing.

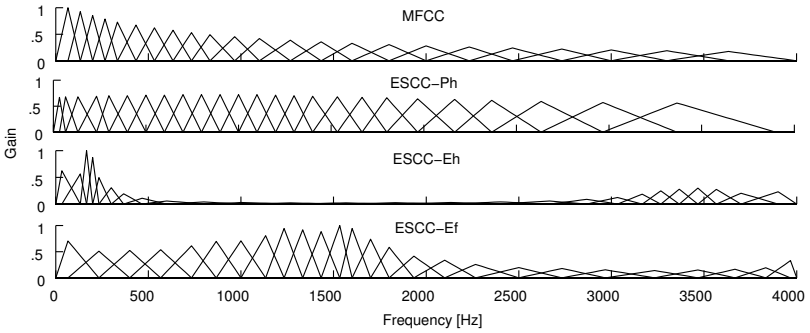

Figure 5: Mel filter bank and optimised filter banks for phoneme recognition, Hindi stressed speech corpus and FAU Aibo emotion corpus, respectively.

[Ooi et al., 2014] Ooi, C. S., Seng, K. P., Ang, L.-M., and Chew, L. W. (2014). A new approach of audio emotion recognition. *Expert Systems with Applications*, 41(13):5858–5869.

[Paul and Das, 2015] Paul, S. and Das, S. (2015). Simultaneous feature selection and weighting – an evolutionary multi-objective optimization approach. *Pattern Recognition Letters*, in press.

[Qin et al., 2013] Qin, Z., Liu, W., and Wan, T. (2013). A bag-of-tones model with MFCC features for musical genre classification. In Motoda, H., Wu, Z., Cao, L., Zaiane, O., Yao, M., and Wang, W., editors, *Advanced Data Mining and Applications*, volume 8346 of *Lecture Notes in Computer Science*, pages 564–575. Springer Berlin Heidelberg.

[Reyes-Vargas et al., 2013] Reyes-Vargas, M., Sánchez-Gutiérrez, M., Rufiner, L., Albornoz, M., Vignolo, L., Martínez-Licona, F., and Goddard-Close, J. (2013). Hierarchical clustering and classification of emotions in human speech using confusion matrices. In *Lecture Notes in Artificial Intelligence*, volume 8113, pages 162–169. Springer.

[Rosenberg, 2012] Rosenberg, A. (2012). Classifying skewed data: Importance weighting to optimize average recall. In *INTERSPEECH 2012*, Portland, USA.

[Sahidullah and Saha, 2013] Sahidullah, M. and Saha, G. (2013). A novel windowing technique for efficient computation of MFCC for speaker recognition. *Signal Processing Letters, IEEE*, 20(2):149–152.

[Sainath et al., 2013] Sainath, T., Kingsbury, B., Mohamed, A.-R., and Ramabhadran, B. (2013). Learning filter banks within a deep neural network framework. In *Automatic Speech Recognition and Understanding (ASRU), 2013 IEEE Workshop on*, pages 297–302.

[Sanchez-Diaz et al., 2014] Sanchez-Diaz, G., Diaz-Sanchez, G., Mora-Gonzalez, M., Piza-Davila, I., Aguirre-Salado, C. A., Huerta-Cuellar, G., Reyes-Cardenas, O., and Cardenas-Tristan, A. (2014). An evolutionary algorithm with acceleration operator to generate a subset of typical testors. *Pattern Recognition Letters*, 41:34–42. Supervised and Unsupervised Classification Techniques and their Applications.

[Sarikaya and Hansen, 2000] Sarikaya, R. and Hansen, J. (2000). High resolution speech feature parametrization for monophone-based stressed speech recognition. *Signal Processing Letters, IEEE*, 7(7):182–185.

[Schuller et al., 2006] Schuller, B., Reiter, S., and Rigoll, G. (2006). Evolutionary feature generation in speech emotion recognition. In *Multimedia and Expo, 2006 IEEE International Conference on*, pages 5–8.

[Schuller et al., 2009] Schuller, B., Steidl, S., and Batliner, A. (2009). The interspeech 2009 emotion challenge. In *INTERSPEECH*, volume 2009, pages 312–315.

[Schuller et al., 2010] Schuller, B., Steidl, S., Batliner, A., Burkhardt, F., Devillers, L., Müller, C. A., and Narayanan, S. S. (2010). The interspeech 2010 paralinguistic challenge. In *INTERSPEECH*, volume 2010, pages 2795–2798. International Speech Communication Association.

[Schuller et al., 2011] Schuller, B., Steidl, S., Batliner, A., Schiel, F., and Krajewski, J. (2011). The interspeech 2011 speaker state challenge. In *INTERSPEECH*, pages 3201–3204. International Speech Communication Association.

[Schuller et al., 2013] Schuller, B., Steidl, S., Batliner, A., Vinciarelli, A., Scherer, K., Ringeval, F., Chetouani, M., Weninger, F., Eyben, F., Marchi, E., et al. (2013). The interspeech 2013 computational paralinguistics challenge: social signals, conflict, emotion, autism. In *INTERSPEECH*, Lyon. International Speech Communication Association.

[Shao et al., 2009] Shao, Y., Jin, Z., Wang, D., and Srinivasan, S. (2009). An auditory-based feature for robust speech recognition. In *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*, pages 4625–4628.

[Shukla et al., 2011] Shukla, S., Prasanna, S., and Dandapat, S. (2011). Stressed speech processing: Human vs automatic in non-professional speakers scenario. In *Communications (NCC), 2011 National Conference on*, pages 1–5.

[Skowronski and Harris, 2004] Skowronski, M. and Harris, J. (2004). Exploiting independent filter bandwidth of human factor cepstral coefficients in automatic speech recognition. *The Journal of the Acoustical Society of America*, 116(3):1774–1780.

[Steidl, 2009] Steidl, S. (2009). *Automatic Classification of Emotion-Related User States in Spontaneous Children's Speech*. Logos Verlag.

[Steinwart and Christmann, 2008] Steinwart, I. and Christmann, A. (2008). *Support vector machines*. Springer.

[Ververidis and Kotropoulos, 2006] Ververidis, D. and Kotropoulos, C. (2006). Emotional speech recognition: Resources, features, and methods. *Speech Communication*, 48(9):1162–1181.

[Vignolo et al., 2013a] Vignolo, L. D., Milone, D. H., and Rufiner, H. L. (2013a). Genetic wavelet packets for speech recognition. *Expert Systems with Applications*, 40(6):2350–2359.

[Vignolo et al., 2013b] Vignolo, L. D., Milone, D. H., and Scharcanski, J. (2013b). Feature selection for face recognition based on multi-objective evolutionary wrappers. *Expert Systems with Applications*, 40(13):5077–5084.

[Vignolo et al., 2011a] Vignolo, L. D., Rufiner, H. L., Milone, D. H., and Goddard, J. C. (2011a). Evolutionary Cepstral Coefficients. *Applied Soft Computing*, 11(4):3419–3428.

[Vignolo et al., 2011b] Vignolo, L. D., Rufiner, H. L., Milone, D. H., and Goddard, J. C. (2011b). Evolutionary Splines for Cepstral Filterbank Optimization in Phoneme Classification. *EURASIP Journal on Advances in Signal Proc.*, 2011:8:1–8:14.

[Wu and Cao, 2005] Wu, Z. and Cao, Z. (2005). Improved MFCC-Based Feature for Robust Speaker Identification. *Tsinghua Science & Technology*, 10(2):158–161.

[Yeganeh et al., 2008] Yeganeh, H., Ahadi, S., Mirrezaie, S., and Ziaei, A. (2008). Weighting of Mel Sub-bands Based on SNR/Entropy for Robust ASR. In *Signal Processing and Information Technology, 2008. ISSPIT 2008. IEEE International Symposium on*, pages 292–296.

[Zão et al., 2014] Zão, L., Cavalcante, D., and Coelho, R. (2014). Time-frequency feature and AMS-GMM mask for acoustic emotion classification. *Signal Processing Letters, IEEE*, PP(99):1–1.

[Zamani et al., 2011] Zamani, B., Akbari, A., Nasersharif, B., and Jalalvand, A. (2011). Optimized discriminative transformations for speech features based on minimum classification error. *Pattern Recognition Letters*, 32(7):948–955.

[Zheng et al., 2001] Zheng, F., Zhang, G., and Song, Z. (2001). Comparison of different implementations of MFCC. *Journal of Computer Science and Tech.*, 16(6):582–589.

[Zheng et al., 2014] Zheng, W., Xin, M., Wang, X., and Wang, B. (2014). A novel speech emotion recognition method via incomplete sparse least square regression. *Signal Processing Letters, IEEE*, PP(99):1–1.

[Zhou et al., 2007] Zhou, X., Fu, Y., Liu, M., Hasegawa-Johnson, M., and Huang, T. (2007). Robust Analysis and Weighting on MFCC Components for Speech Recognition and Speaker Identification. In *Multimedia and Expo, 2007 IEEE International Conference on*, pages 188–191.