Available online at www.sciencedirect.com

SciVerse ScienceDirect

Computer Speech and Language 28 (2014) 278–294

COMPUTER SPEECH AND LANGUAGE

www.elsevier.com/locate/csl

Shape-based modeling of the fundamental frequency contour for emotion detection in speech

Juan Pablo Arias[a], Carlos Busso[b], Nestor Becerra Yoma[a,*]

[a] *Speech Processing and Transmission Laboratory, Department of Electrical Engineering, Universidad de Chile, Santiago, Chile*
[b] *Multimodal Signal Processing Laboratory, The University of Texas at Dallas, Richardson, TX 75080, USA*

Received 20 July 2012; received in revised form 25 March 2013; accepted 16 July 2013
Available online 26 July 2013

Abstract

This paper proposes the use of neutral reference models to detect local emotional prominence in the fundamental frequency. A novel approach based on *functional data analysis* (FDA) is presented, which aims to capture the intrinsic variability of F0 contours. The neutral models are represented by a basis of functions and the testing F0 contour is characterized by the projections onto that basis. For a given F0 contour, we estimate the functional *principal component analysis* (PCA) projections, which are used as features for emotion detection. The approach is evaluated with lexicon-dependent (i.e., one functional PCA basis per sentence) and lexicon-independent (i.e., a single functional PCA basis across sentences) models. The experimental results show that the proposed system can lead to accuracies as high as 75.8% in binary emotion classification, which is 6.2% higher than the accuracy achieved by a benchmark system trained with global F0 statistics. The approach can be implemented at sub-sentence level (e.g., 0.5 s segments), facilitating the detection of localized emotional information conveyed within the sentence. The approach is validated with the SEMAINE database, which is a spontaneous corpus. The results indicate that the proposed scheme can be effectively employed in real applications to detect emotional speech.

Keywords: Emotion detection; F0 contour modeling; Emotional speech analysis; Expressive speech

1. Introduction

Emotional understanding is a crucial skill in human communication. It plays an important role not only in interpersonal interactions, but also in many cognitive activities such as rational decision making, perception and learning (Picard, 1997). For this reason, modeling and recognizing emotions is essential in the design and implementation of *human-machine interfaces* (HMIs) that are more in tune with the user's needs. Systems that are aware of the user's emotional state will facilitate several new scientific avenues that serve as truly innovative advancements in security and defense (e.g. threat detection), health informatics (e.g., depression, autism), and education (e.g., tutoring system) (Burleson and Picard, 2004; Langenecker et al., 2005). Given the important role of speech in the expression of emotions, an increasing number of publications have reported progress in automatic emotion recognition and detection using acoustic features. Complete reviews are given by Cowie et al. (2001), Zeng et al. (2009), Schuller et al. (2011a), Koolagudi and Rao (2012), El Ayadi et al. (2011).

[*] Corresponding author. Tel.: +56 2 29784205; fax: +56 2 26953881.
E-mail addresses: nbecerra@ing.uchile.cl, nbecerray@gmail.com (N.B. Yoma).

0885-2308/$ – see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.csl.2013.07.002

The dominant approach in emotion recognition from speech consists in estimating global statistics or functionals at sentence level from low level descriptors such as F0, energy and *Mel-frequency cepstral coefficients* (MFCCs) (Schuller et al., 2011a). Among prosodic based features, gross pitch statistics such as mean, maximum, minimum and range are considered as the most emotionally prominent parameters (Busso et al., 2009). One limitation of global statistics is the assumption that every frame in the sentence is equally important. Studies have shown that emotional information is not uniformly distributed in time (Lee et al., 2004; Busso and Narayanan, 2007). For example, the intonation in happy speech tends to increase at the end of the sentence (Wang et al., 2005). Since the statistics are computed at the global level, it is not possible to identify local salient segments or focal points within the sentence. Furthermore, features describing global statistics do not capture local variations (e.g., in F0 contours), which in turn could provide useful information for emotion detection. In this context, this paper proposes a novel shape-based approach to detect emotionally salient temporal segments in the speech using *functional data analysis* (FDA). The detection of localized emotional segments can shift current approaches in affective computing. Instead of recognizing the emotional content on pre-segmented sentences, the problem can be formulated as a detection paradigm, which is appealing from an application perspective (e.g., continuous assessments of unsegmented recordings). The emotion recognition system can be more robust by weighting each frame according to their emotional saliency. From a speech production viewpoint, the approach can shed light into the underlying interplay between lexical and affective human communication across various acoustic features (Busso and Narayanan, 2007).

This study focuses on detecting emotionally salient temporal segments on the fundamental frequency. Patterson and Ladd (1172) argued that the range (i.e., the difference between the maximum and the minimum of F0 contour in a sentence or utterance) does not give information about the distribution of F0 and hence valuable emotional information is neglected. Also, according to Lieberman and Michaels (1962) low variations in F0 can be subjectively relevant in the identification of emotions. In the literature, there are some attempts to model the shape of the F0 contour. Paeschke and Sendlmeier (2000) analyzed the rising and falling movements of F0 within accents in affective speech. The study incorporated metrics related to accent peaks within a sentence. The authors found that those metrics present statistically significant differences between emotional classes. Also, Paeschke (2004) modeled the global trend of F0 in emotional speech as the gradient of linear regression. The author concluded that global trend can be useful to describe emotions such as boredom and sadness. Rotaru and Litman (2005) employed linear and quadratic regression coefficients and regression error as features to represent pitch curves. Yang and Campbell (2001) argued that concavity and convexity of the F0 contour reflect the underlying expressive state. The *Tone and Break Indices* system (ToBI) is a scheme for labeling prosody that has been widely used for transcribing intonation (Silverman et al., 1992). Liscombe et al. (2003) analyzed affective speech with acoustic features by using ToBI labels to identify the type of nuclear pitch accent, the contour type and the phrase boundaries. Despite the fact that ToBI provides an interesting approach to describe F0 contours, more precise labeling is required to generate prosodic transcripts. Taylor (2000) introduced the *Tilt Intonation Model* to represent intonation as a linear sequence of events (e.g. pitch accents or boundaries), which in turn are given by a set of parameters. However, an automatic event segmentation algorithm is required to employ this scheme and, hence, it cannot be easily applied to emotion recognition or detection tasks.

Despite current efforts to address the problem of affective speech characterization by means of modeling F0 contour, this is still an open task. The contributions of the paper concern: (a) a novel framework to detect emotional modulation based on reference templates that models F0 contours of neutral speech; (b) an insightful and thorough analysis of neutral references as a method to detect emotion in speech; (c) the generation of reference F0 contour templates with *functional data analysis* (FDA); and, (d) a study of the shortest segmentation unit that can be used in emotion detection. Extensive experiments are presented to demonstrate the discriminative power of the FDA based approach to detect emotional speech. The results on the SEMAINE database reveal that the approach captures localized emotional information conveyed in short speech segments (0.5 s). These properties of the proposed approach are interesting from the research and application points of view.

2. Emotional databases and features

2.1. Emotional databases

The analysis and results presented in Sections 3 and 4 require recordings with controlled, lexicon-dependent conditions (e.g., recordings of sentences with the same lexical content conveying different emotional states). Therefore,

Table 1
Databases (N = neutral, A = anger, H = happiness, S = sadness, F = fear, D = disgust, B = boredom, Va = valence, Ac = activation/arousal, Po = power, Ex = anticipation/expectation, In = intensity). Number of samples per emotion is given in brackets.

Database	WSJ1	EMA	EMO-DB	SEMAINE
Type	Neutral	Emotional	Emotional	Emotional
Use of data	Reference	Train/test	Train/test	Test
Spontaneous/acted	Spon.	Acted	Acted	Spon.
# Speakers	50	3	10	10
# Utterances	8104	680	535	1926

Emotions/attributes	N(8104)	N(170),A(170)
		H(170),S(170)
			N(79),F(69),D(46),H(71)
			B(81),S(62),A(127)
				Va, Ar, Po
				Ex, In

the study considers, for these sections, two emotional databases recorded from actors (Table 1). Even though acted emotions differ from real-life emotional manifestation, they provide a good first approximation, especially when controlled conditions are required, as in this study. The first database is the EMA corpus collected at the *University of Southern California* (USC) (Lee et al., 2005).[1] Three speakers participated in the recordings (two of them with formal theatrical vocal training). They read ten sentences five times in happy, angry, sad and neutral conditions (one subject read 4 additional sentences producing 80 extra samples). The subjects were asked to record the sentences in random order to attenuate or eliminate reproductions with similar intonation. To reduce fatigue, the recording was split into small sessions separated by breaks. The EMA database was evaluated by four native speakers of American English in terms of the emotional classes happy, angry, sad, neutral and other. The average human recognition rate was 81.8% (Grimm et al., 2007). The second corpus is the *Berlin Database of Emotional Speech* (EMO-DB) (Burkhardt et al., 2005). This database is composed of ten speaker (five male and five female), who read ten German sentences one time expressing six different emotions (fear, disgust, happiness, boredom, sadness, anger), in addition to neutral state. This database is available to the community and has been widely used in related work on emotion recognition. Therefore, other researchers can easily replicate the results presented here.

After relaxing the controlled lexicon-dependent conditions, the framework is validated using a spontaneous emotional corpus (Section 5). The study considers the SEMAINE database, which includes audiovisual recordings of natural human computer interactions (McKeown et al., 2010). The emotions are elicited using the *Sensitive Artificial Listener* (SAL) approach. We consider sessions recorded from ten subjects. The data contains subjective evaluations generated by human raters using Feeltrace (Cowie et al., 2000). This is a labeling tool employed to continuously track the perceived emotional state over time (as oppose to assigning one discrete label per sentence). The raters are asked to move the cursor as they watch/listen the stimulus using a *graphical user interface* (GUI). The GUI records the position of the pointer, which describes the emotional content in term of continuous attributes. Although the corpus has been annotated with various emotional attributes, we consider only the activation/arousal (calm versus active) and valence (negative versus positive) dimensions.

2.2. F0 extraction and post-processing

The fundamental frequency is estimated according to the following steps: first, speech signals are divided into 400-sample (25 ms) frames with 50% overlap. The fundamental frequency is estimated by using the autocorrelation based Praat pitch detector system (Boersma and Weenink, 1996). Then, F0 at each frame is represented in a semitone scale according to:

$$F0_{semitone}(t) = 12\frac{\log\left[F0(t)\right]}{\log 2} \tag{1}$$

where $F0(t)$ and $F0_{semitone}(t)$ are the fundamental frequency at frame t in Hertz and semitones, respectively. The proposed scheme in this paper aims to model the F0 contour to compare neutral and emotional speech. Consequently, the logarithm attempts to represent differences in F0 according to the human-like perception scale. After estimating

[1] The EMA database is available at http://sail.usc.edu/ema_web/

$F0_{semitone}(t)$, unvoiced segments are interpolated with cubic spline to obtain smooth and continuous F0 contours. Finally, the resulting interpolated $F0_{semitone}(t)$ contour is normalized by subtracting the mean. Henceforth, the term “F0 contour” denotes the F0 curve in the semitone scale after interpolation and mean normalization.

3. Proposed method

Building neutral reference models to contrast emotional speech is an appealing method. The scheme significantly reduces the dependency on emotional (acted or spontaneous) speech databases, which in turn are much more difficult to obtain than ordinary corpora. This section describes the proposed neutral reference models built with FDA.

3.1. Motivation

First, we present an experiment showing that a single neutral speech signal can be used as a reference to detect emotional prominence in the fundamental frequency. We compare the F0 contour extracted from an testing signal (either emotional or neutral speech) with the one extracted from a neutral reference sentence conveying the same lexical information. Notice that this case corresponds to the ideal scenario where the testing and reference utterances convey the same verbal information. To keep all variables under control except emotional modulation, the analysis takes place with speaker matching condition and with the same lexical content in both testing and reference utterances. Given these constrains, this experiment is conducted using only the EMA database (e.g., the subjects produced five repetitions for each sentence and each emotion).

The reference and testing sentences are compared using a top-down strategy similar to our previous work (Arias et al., 2010). The F0 extraction and post-processing steps are applied to both utterances. Since the reference and testing utterances are not temporally aligned, they are aligned according to their MFCCs by using standard *dynamic time warping* (DTW) (Euclidian distance, slope constraint condition $P=0$ and the Sakoe–Chiba band as global path restriction). After aligning the signals, the Pearson’s correlation is employed as a similarity measure to estimate the differences between both F0 patterns. Lower correlation levels will indicate higher differences between neutral and emotional sentences, which in turn can be mainly associated to emotional modulation in the testing utterance. Given a speaker, a neutral sentence is compared with its emotional versions using all possible permutations (i.e., happy, angry and sad). Likewise, we compare all possible permutation of neutral-neutral sentences (the reference signal is always neutral). Fig. 1 presents the distribution of the correlation based similarity measure between testing and reference utterances for each emotion. According to Fig. 1 the comparison of neutral testing with reference patterns, which in turn are neutral by definition, provides the highest correlation based similarities ($\overline{\rho} = 0.84 \pm 0.15$). This result indicates that emotionally neutral utterances with the same lexical content spoken by the same speaker produce similar F0 contours. In contrast, the similarity between F0 contours provided by emotional speech and neutral reference utterances is significantly lower. The results of this experiment suggest that the similarity of F0 contours extracted from the testing and reference utterances can be used to detect the emotional state in speaker and lexicon matched conditions. This paper explores a shape-based approach using FDA to generalize these ideas for unconstrained scenarios (speaker-independent, lexicon-independent approach).

3.2. Functional data analysis (FDA)

The idea behind FDA is to represent the structure of signals as functions by using statistical methods (Ramsay and Silverman, 2005). The time series data is represented by a continuous and smooth function –$x(t)$– that is generated as a linear combination of basis functions ϕ_k such as B-spline or polynomial:

$$x(t) = \sum_{k=1}^{K} c_k \phi_k(t). \tag{2}$$

where K represents the dimension of the expansion and c_k corresponds to the projection onto the kth basis function. Both ϕ_k and K are parameters of FDA that should be properly chosen according to the characteristics of the data.

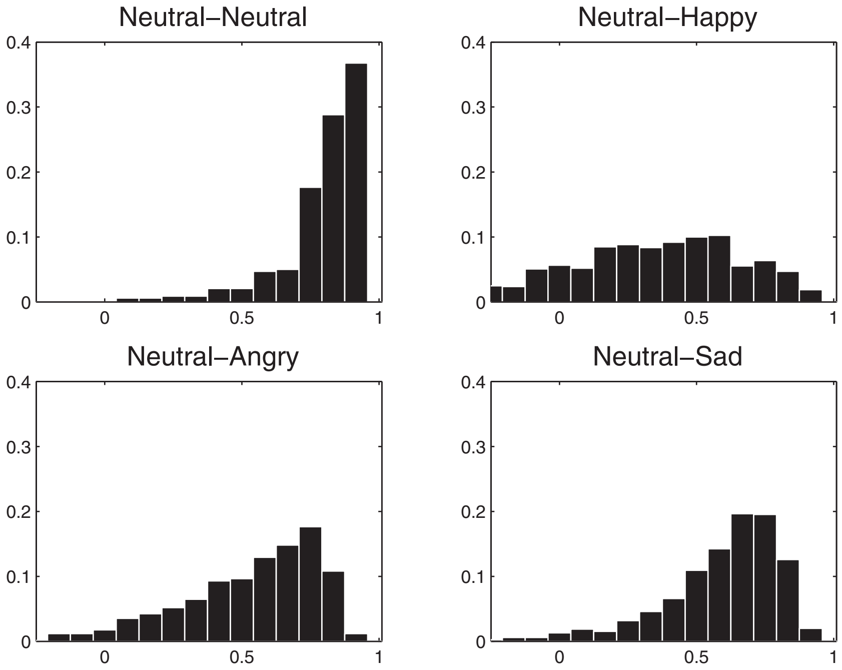

Fig. 1. Similarity measure distributions for neutral, angry, happy and sad samples (EMA). The figure shows the Pearson's correlation between neutral and emotional F0 contours.

Functional data is observed as a discrete sequence $(t_j, y_j), j \in \{1, \ldots, n\}$, where y_j corresponds to the sampled value of the function $x(t)$ at time t_j. This sequence may not be equally-distributed and may be corrupted by noise, ϵ_j:

$$y_j = x(t_j) + \epsilon_j. \tag{3}$$

In FDA, the process of fitting functions to data is known as *smoothing*. Given discrete observations y_j and basis function $\{\phi_1, \ldots, \phi_K\}$, this process attempts to find coefficients c_k by minimizing the mean squared error ϵ_j. A roughness penalty term is incorporated in the optimization to ensure a smooth representation. Optimal parameters $\hat{c_k}$ are estimated with Eq. (4), where λ is a smoothing parameter and D^m represents the mth derivative (Ramsay and Silverman, 2005).

$$\hat{c_k} = \underset{c_k}{\operatorname{argmin}} \sum_{j=1}^{n} [y_j - x(t_j)]^2 + \lambda \int [D^m x(s)]^2 ds \tag{4}$$

FDA provides several advantages when compared with classical approaches that represent data as a set of isolated samples. Descriptive statistics such as mean, covariance and correlation can be applied to functional data. These properties are useful to model and analyze F0 contours. More interestingly in the context of this paper, powerful tools for analyzing data such as *principal component analysis* (PCA) can be employed in the framework of FDA. Conventional PCA is a method that converts a set of correlated observations into uncorrelated variables, called *principal components* (PCs), by using orthogonal transformations. Functional PCA is a technique that extends this framework to the domain of functions (Ramsay and Silverman, 2005). Given a set of V functions, denoted by $x_v(t)$, the *principal components scores*, $f_{u,v}$, are given by

$$f_{u,v} = \int \xi_u(t) x_v(t) dt \tag{5}$$

where $\xi_u(t)$ corresponds to an orthonormal basis denoted as *principal component functions*, that represents the variability of $x_v(t)$. The first PC function $\xi_1(t)$ is estimated by maximizing $\sum_v f_{1,v}^2$, subject to the constrain $\int \xi_1(t)^2 dt = 1$. Similarly, the subsequent function vectors $\xi_u(t)$ are obtained by maximizing $\sum_v f_{u,v}^2$ subject to the constrain $\int \xi_u P(t)^2 dt = 1$ and

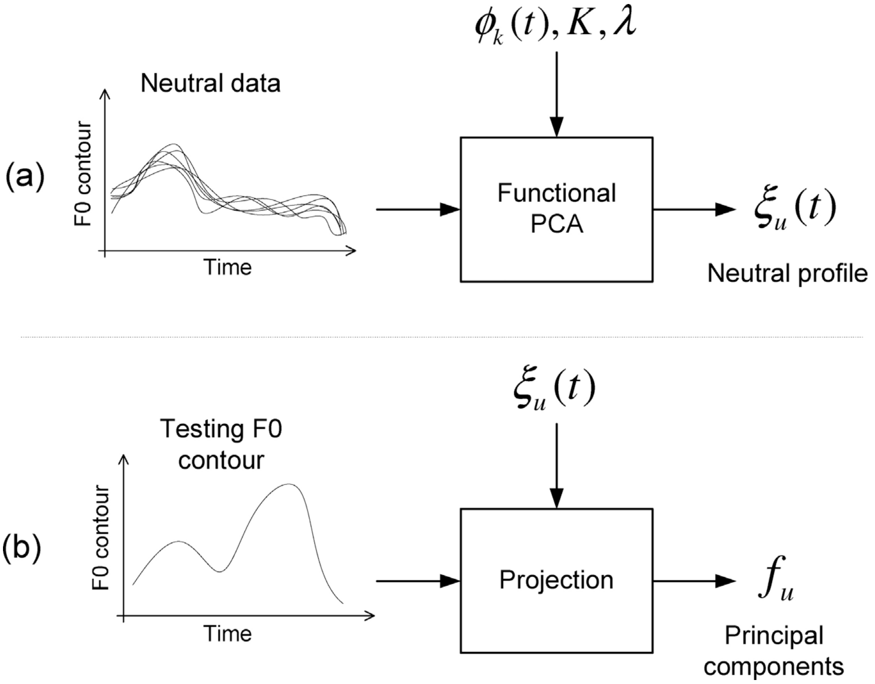

Fig. 2. General framework of the proposed approach: (a) neutral model generation by using Functional PCA and (b) projection of a testing utterance onto the neutral eigenspace.

to the additional $m-1$ constrain(s) $\int \xi_u(t)\xi_m(t)dt = 0$, $\forall m < u$. Finally, functions $x_v(t)$ can be approximated by using the first U principal components:

$$\hat{x}_v(t) = \sum_{u=1}^{U} f_{u,v}\xi_u(t). \tag{6}$$

FDA provides an interesting framework to model F0 contours (Gubian et al., 2009; Zellers et al., 2010). In particular, functional PCA allows us to statistically represent a family of functions, which can be employed as neutral reference to contrast emotional speech. It is worth mentioning that FDA has been employed to provide a descriptive analysis of prosodic features in previous research (Gubian et al., 2009; Zellers et al., 2010). In contrast, this paper proposes to apply FDA as a tool for generating neutral models and projections to represent F0 contours from a pattern recognition point of view.

3.3. Functional PCA reference models for F0 contours

Fig. 2(a) shows the general framework to build neutral references with functional PCA. First, a set of neutral utterances spoken by several speakers are employed as training data. All the utterances are temporally aligned with standard DTW. Then, the F0 contour extraction procedure described in Section 2.2 is applied to the signals. The resulting time-aligned, post-processed F0 contours are smoothed and represented as functional data by employing a basis of B-spline functions $\phi_k(t)$ according to Eqs. (2) and (4). Finally, functional PCA is applied to generate a new orthogonal basis of functions $\xi_u(t)$.

Fig. 2(b) shows the testing stage of the proposed scheme. As a first step, the testing utterance is aligned with the training data using DTW. Then, F0 contour is extracted and the projections of the testing F0 contour onto the neutral reference basis $\xi_u(t)$ is estimated. As a result, the coefficients f_u are obtained, which correspond to the parameters that describe the shape of the testing F0 contour. Since the profile $\xi_u(t)$ is generated with non-emotional speech, it is expected that neutral and emotional testing F0 contours will provide different projections (i.e. $\{f_1 \ldots f_U\}$) onto the functional PCA basis. Therefore, the set of parameters $\{f_1 \ldots f_U\}$ could be used to detect emotional speech.

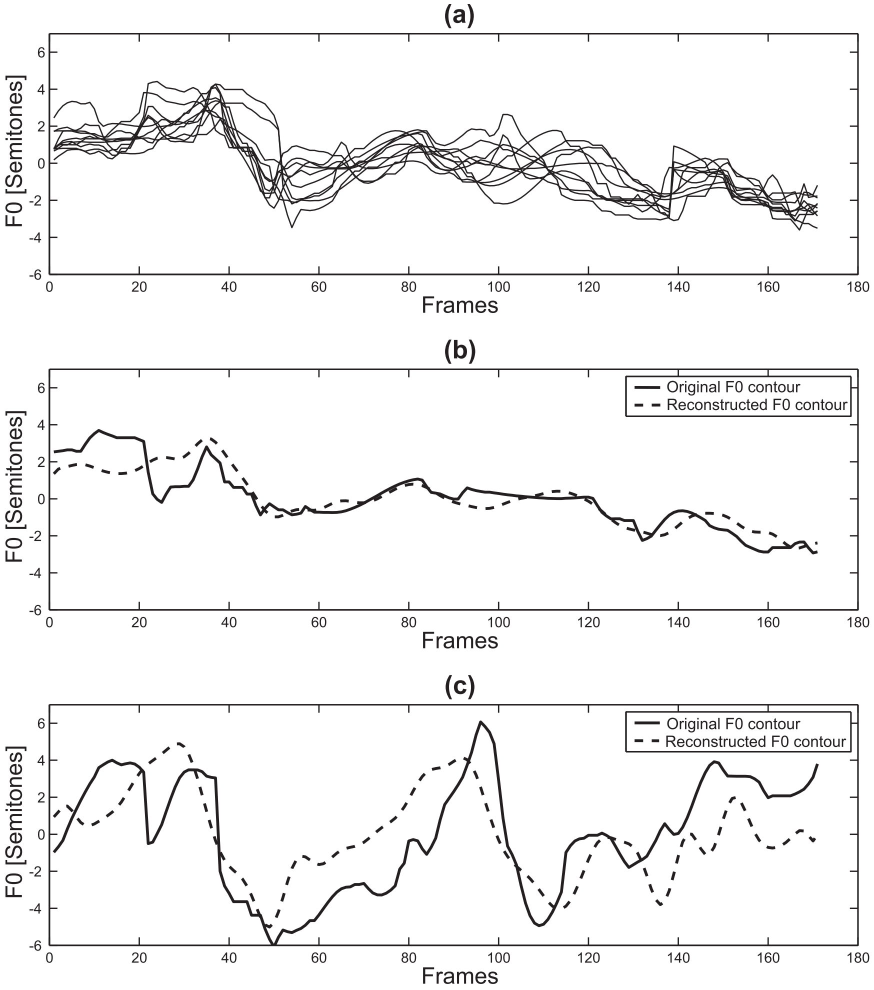

Fig. 3. Reconstruction of F0 contours with Functional PCA: (a) training data to generate the neutral functional PCA basis; (b) reconstruction of a neutral testing utterance with the first five principal components; and (c) reconstruction of a happy testing utterance with the first five principal components. The mean squared error between the original and the reconstructed F0 contours is equal to 0.45 and 0.32 for the neutral and happy utterances.

As an illustration, we implement the approach with lexicon-dependent models – a FDA reference model for a given sentence. Fig. 3 presents an example for the sentence “*I am talking about the same picture you showed me*” extracted from the EMA database. Fig. 3(a) shows the time-aligned and post-processed F0 curves of ten neutral realizations uttered by two speakers (five repetitions each one). Although the sentences present variations in their F0 contours, they clearly have a pattern that our approach aims to capture. This result agrees with previous studies that have shown that matching the linguistic content of the testing sentence improves the emotion classification accuracy (Vlasenko et al., 2008). Then, a neutral profile is trained with this data by applying the procedure presented in Fig. 2(a). For this example, the smoothing basis ϕ_k is implemented with a 6th order B-spline function basis with $K = 40$. Fig. 3(b) and (c) show the reconstruction of neutral and happy F0 contours, respectively, for the same sentence, uttered by a third subject whose data was not considered to build the neutral reference. Both F0 curves are reconstructed using the first five principal components. As can be seen in Fig. 3(b) and (c), the neutral F0 contour is more accurately approximated with the neutral functional PCA basis than the F0 contour corresponding to happy speech. The reference model fits better the neutral sentence than the one corresponding to happy speech. Therefore, it is reasonable to conclude that the projection onto the kth basis function, with $6 \leq k \leq 40$, converges to zero faster with the neutral utterance than with the happy sentence. Consequently, this analysis strongly suggests that the projections of F0 contours from emotional speech are different from those generated from neutral utterances. This result is supported by Fig. 4 that shows the

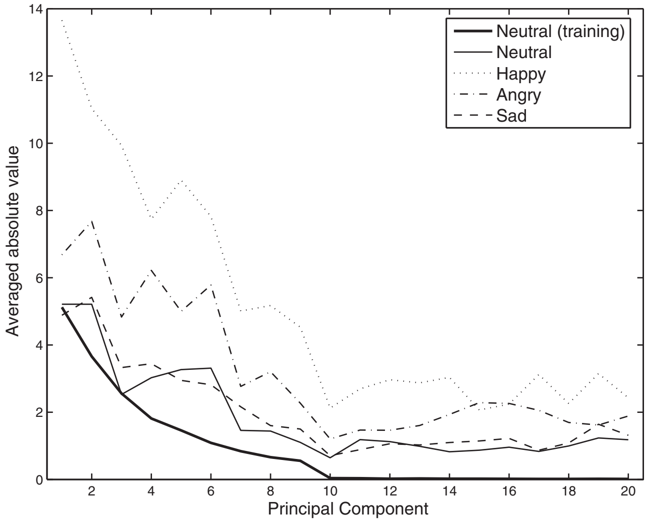

Fig. 4. Averaged absolute value of the projections associated to each principal component (EMA database). The figure shows that functional PCA projections can be used as features.

averaged absolute value of the projections onto the first 20 principal components for neutral and emotional speech. The projections are generated by using the leave-one-speaker-out strategy, in which the functional PCA models are trained with two speakers and tested with the third one in the EMA corpus. In contrast to emotional speech, Fig. 4 shows that the mean of the projections of the F0 contours from neutral speech are approximately equal to zero when $k \geq 10$. The averaged absolute value of the projections for happy and angry speech are higher than the neutral ones, even for higher order principal components.

4. Discriminant analysis

To assess the discriminative power of the functional PCA projections, this section evaluates the approach using lexicon-dependent models (e.g., one functional PCA model for each utterance – Section 4.1), and a lexicon-independent model (e.g., a single functional PCA model for all sentences – Section 4.2). It also evaluates the performance of the approach at sub-sentence level (Section 4.3). The evaluation considers both the EMA and EMO-DB databases, separately.

The approach is evaluated for different emotion detection problems, in which we implement binary emotion classification between neutral and emotional speech. We train a separate binary classifier for neutral speech and each of the emotional categories in the EMA and EMO-DB databases (e.g., neutral-happiness, neutral-anger, and neutral-sadness). In addition, an extra class, denoted as “emotional”, is generated for each corpus by grouping utterances from all the emotional classes. Given the aggregation of emotional classes, there are more samples in the emotional class. Therefore, for these neutral-emotional tasks, we randomly choose emotional samples to match the number of neutral samples (chance = 50%). This procedure is repeated 100 times and the performance rates are averaged. This approach is also implemented with the EMO-DB corpus, since the emotional classes are unbalanced (see Table 1).

All the binary classification experiments are implemented with a *quadratic discriminant classifier* (QDC). QDC estimates the output score by using a quadratic combination of the feature vector: $y = x^T ax + b^T x + c$, where x and y are the input vector and the output score, respectively. Notice that our preliminary experiments show that QDC provides similar performance than a *support vector machine* (SVM), which has been widely used in previous emotion recognition problems (Schuller et al., 2010). While the performance of a SVM depends on the kernel and the soft margin parameter, which have to be estimated over a validation partition, all the parameters of QDC are easily estimated from the training set. Therefore, QDC is chosen for its simplicity, performance, consistency and generalization.

For comparison purposes, a benchmark system was implemented with QDC using a conventional approach, in which the classifiers are trained with statistics derived from the F0 contour (Busso et al., 2009, 2013; Zeng et al., 2009; Cowie et al., 2001). First, 80 sentence-level functionals derived from F0 contour were extracted for each utterance by

Table 2
Benchmark system (EMA and EMO-DB). Features were extracted from F0 contour at sentence level (Acc = accuracy, Pre = precision, Rec = recall, *F* = *F*-score). Chance is the ratio between number of emotional samples and total number of utterances.

		Acc	Pre	Rec	*F*	Chance
EMA	Neutral-happy	0.843 (0.086)	0.927	0.780	0.819	0.500
	Neutral-angry	0.737 (0.120)	0.882	0.627	0.666	0.500
	Neutral-sadness	0.653 (0.195)	0.683	0.753	0.687	0.500
	Neutral-emotional	0.714 (0.116)	0.723	0.762	0.732	0.500
EMO-DB	Neutral-fear	0.642 (0.121)	0.861	0.370	0.469	0.500
	Neutral-disgust	0.658 (0.108)	0.726	0.562	0.583	0.500
	Neutral-happiness	0.763 (0.094)	0.964	0.552	0.679	0.500
	Neutral-boredom	0.513 (0.016)	0.875	0.055	0.102	0.500
	Neutral-sadness	0.706 (0.087)	0.644	0.966	0.769	0.500
	Neutral-anger	0.825 (0.120)	0.930	0.709	0.774	0.500
	Neutral-emotional	0.690 (0.097)	0.889	0.458	0.555	0.500

using the openSMILE audio feature extraction toolkit (Eyben et al., 2009). The set of functionals corresponds to those employed for the Interspeech 2010 Paralinguistic Challenge (Schuller et al., 2010). Then, *forward feature selection* (FFS) was applied to reduce the number of features to 20, matching the number of projections used as features in the proposed approach. For the EMA database, the classifiers are trained with two speakers and tested with a third one. Three permutations were implemented by interchanging the role of each speaker between training and testing data sets. For the EMO-DB database, the corpus was also split in speaker independent partitions, in which all the speech samples from one subject are included in either training or testing sets. We also use a cross-validation approach and the average results are reported in Table 2.

4.1. Evaluation with lexicon-dependent functional PCA models

First, we evaluate the proposed approach by building a functional PCA reference model for each sentence in the corpora. The EMA database was divided in development (to build the functional PCA reference models), training (to train the classifier) and testing (to estimate the accuracy) sets. Each of these three sets contains speech samples from only a single speaker. Only neutral data was employed to build the reference models. To maximize the use of the EMA database, six permutations were implemented by interchanging the role of each speaker among development, training and testing data sets. This procedure ensures that the results are speaker-independent. The performance rates were estimated by averaging the results obtained in all six implementations. A similar procedure was implemented for the EMO-DB database. The sentences were partitioned into development, training and testing subsets.

Table 3
Discriminant analysis with functional PCA projections using lexicon-dependent bases (EMA, EMO-DB) (Acc = accuracy, Pre = precision, Rec = recall, *F* = *F*-score). Chance is the ratio between number of emotional samples and total number of utterances.

		Acc	Pre	Rec	*F*	Chance
EMA	Neutral-happy	0.913 (0.033)	0.884	0.960	0.918	0.500
	Neutral-angry	0.777 (0.071)	0.784	0.780	0.777	0.500
	Neutral-sadness	0.633 (0.067)	0.626	0.693	0.654	0.500
	Neutral-emotional	0.758 (0.057)	0.785	0.726	0.752	0.500
EMO-DB	Neutral-fear	0.709 (0.039)	0.715	0.706	0.707	0.500
	Neutral-disgust	0.710 (0.039)	0.715	0.707	0.707	0.500
	Neutral-happiness	0.736 (0.025)	0.742	0.725	0.733	0.500
	Neutral-boredom	0.639 (0.051)	0.661	0.562	0.604	0.500
	Neutral-sadness	0.680 (0.035)	0.718	0.598	0.646	0.500
	Neutral-anger	0.738 (0.013)	0.738	0.740	0.738	0.500
	Neutral-emotional	0.699 (0.011)	0.715	0.663	0.687	0.500

Table 4
Proportion hypothesis test to compare classifiers (EMA). Light and dark gray values represent strong (p-value<0.05) and weak (p-value<0.1) statistic significance, respectively (B = benchmark, LD = lexicon-dependent models, LI = lexicon-independent models, TB = time-based segmentation, Ch = chunk-level segmentation, Wo = word-level segmentation).

	Happiness	Anger	Sadness	Emotion
LD - B	0.001	0.092	0.278	0.078
LI - B	0.001	0.088	0.283	0.081
LD - LI	0.121	0.219	0.001	0.266
LI - TB	0.228	0.431	0.425	0.304
LI - Ch	0.041	0.431	0.030	0.001
LI - Wo	0.001	0.018	0.015	0.190

Table 3 shows the performance of the proposed system at the sentence-level. For the EMA database, the accuracy for neutral-happiness task is 91.3%. Also, the accuracies for both neutral-anger and neutral-emotional tasks are higher than 75%. These results strongly validate the proposed scheme. For the EMO-DB database, the accuracies for neutral-happy and neutral-anger classification tasks are higher than 73%. As expected, the accuracy for the neutral-sadness task is low for both databases (EMA 63.3%, EMO-DB 68%). These results are consistent with the analysis presented in Figs. 1 and 4 that shows that the discrimination between neutral and sad classes is lower than in the neutral-happiness and neutral-anger cases.

In general, the accuracies achieved with the proposed approach (Table 3) are higher than the ones achieved by the benchmark classifier (Table 2). In the EMA database, the accuracy of the proposed system in the neutral-happiness task is 7.0% (absolute) higher than the benchmark system (statistically significant with p-value = 0.001, see Table 4). Also, the classifiers neutral-anger and neutral-emotional achieves 4.0% and 4.4% (absolute) improvements, respectively, when compared with the benchmark classifier (p-value = 0.092 and p-value = 0.078, respectively, see Table 4). In the EMO-DB database, the proposed method leads to an increase in accuracy for neutral-fear, neutral-disgust and neutral-boredom tasks equal to 6.8%, 5.2% and 12.6% (absolute), respectively. These results suggest that the proposed scheme can accurately discriminate between neutral and emotional categories. However, the accuracy achieved by the functional PCA based system is lower for the neutral-sadness task for the EMA (2%) and EMO-DB (2.6%) databases. It is worth highlighting that, when compared with the benchmark method, the improvements in accuracy achieved by the proposed system in the neutral-emotional, neutral-happiness and neutral-anger classification tasks are much higher than the accuracy degradation with the neutral-sadness classification. Moreover, the standard deviations of the accuracy obtained by the proposed system (Table 3) are much lower than those with the benchmark method (Table 2). These results strongly suggest that the functional PCA based classifier is more reliable and consistent than the benchmark method based on F0 statistics.

4.2. Evaluation with lexicon-independent functional PCA models

The results presented in Section 4.1 reveal that the FDA-based approach proposed in this paper can accurately discriminate between neutral and emotional speech. However, a lexicon-dependent emotion classifier system loses applicability in real applications. This section evaluates the proposed approach with lexicon-independent models. Basically, the idea is to build the functional PCA basis with neutral sentences conveying different lexical information. The results in Section 3.3 suggested that lexical information affects the F0 contour even for non-tonal languages. Lexicon-independent bases will not capture this aspect. However, by relaxing the constraint of using sentences with the same verbal message, more sentences can be used to build the functional PCA basis. This approach will produce more robust reference models that will better capture the F0 variability.

The lexicon-independent F0 contours are extracted and post processed from the neutral sentences according to the approach described in Section 2.2. Then, the average duration of the signals is estimated and used to linearly warp their F0 contours. The resulting family of neutral F0 contours are used as input for functional PCA (Fig. 2(a)). In order to evaluate the performance of the lexicon-independent system, the same discriminant analysis described in Section 4.1 was followed. Observe that for those experiments there were 10 speaker-dependent neutral functional PCA models per corpus (i.e., one per each sentence). In this section, there is only one text-independent functional PCA model trained with all the neutral F0 contours.

Table 5
Discriminant analysis with functional PCA projections using lexicon-independent bases (EMA, EMO-DB) (Acc = accuracy, Pre = precision, Rec = recall, F = F-score). Chance is the ratio between number of emotional samples and total number of utterances.

		Acc	Pre	Rec	F	Chance
EMA	Neutral-happy	0.893 (0.059)	0.874	0.933	0.899	0.500
	Neutral-angry	0.795 (0.059)	0.793	0.827	0.802	0.500
	Neutral-sadness	0.547 (0.063)	0.550	0.573	0.553	0.500
	Neutral-emotional	0.742 (0.057)	0.784	0.701	0.733	0.500
EMO-DB	Neutral-fear	0.709 (0.050)	0.737	0.673	0.685	0.500
	Neutral-disgust	0.711 (0.049)	0.738	0.675	0.686	0.500
	Neutral-happiness	0.789 (0.032)	0.788	0.806	0.793	0.500
	Neutral-boredom	0.706 (0.057)	0.753	0.614	0.674	0.500
	Neutral-sadness	0.663 (0.056)	0.807	0.432	0.557	0.500
	Neutral-anger	0.777 (0.034)	0.780	0.785	0.779	0.500
	Neutral-emotional	0.713 (0.036)	0.756	0.641	0.691	0.500

Table 5 presents the classification results with functional PCA projections by using lexicon-independent models with both EMA and EMO-DB databases. For the EMA database, the accuracies in lexicon-independent models for neutral-happiness and neutral-emotional tasks are only 2.0% (absolute) and 1.6% (absolute) lower than the ones in lexicon-dependent models (see Table 3 and 5). For the EMO-DB database, the accuracy of the neutral-emotional classification task achieved with lexicon-independent models is 1.5% (absolute) higher than the accuracy achieved with lexicon-dependent models. According to a proportion hypothesis test, these differences are not statistically significant (see Table 4). Moreover, when compared with the benchmark system (Table 2), the proposed lexicon-independent functional PCA based system leads to absolute improvements in accuracy equal to 5.0%, 5.8% and 2.8% for the neutral-happiness, neutral-anger and neutral-emotional classification tasks, respectively (EMA database). All these differences are statistically significant (see Table 4). Similarly with the EMO-DB database, when compared with the benchmark system (Table 2), the proposed lexicon-independent system leads to improvements in accuracy equal to 6.7%, 5.2% and 2.3% for the neutral-fear, neutral-disgust and neutral-emotional tasks.

4.3. *Evaluation of emotional prominence at sub-sentence level*

This section evaluates the proposed lexicon-independent functional PCA based method implemented at sub-sentence levels (e.g. chunk, word). Emotional prominence conveyed in the F0 contour is not uniformly distributed in time (Busso et al., 2009; Busso and Narayanan, 2007; Yildirim et al., 2004). By extending the analysis to sub-sentence units, we aim to detect these emotionally salient segments. This approach does not require to split a dialog into sentences. Therefore, it is applicable for real time emotion detection systems.

The functional PCA-based emotion detection system shown in Fig. 2 is applied to three different sub-sentence units: time-based segment; phrase (or chunk); and, word. Time-based segmentation consists in dividing the speech signal into one-second windows with 50% of overlap (Jeon et al., 2011). This segmentation does not require to estimate syntactic boundaries. A phrase or chunk is defined as a group of words that form a constituent working as a syntax single unit, which in turn is suitable for emotion recognition (Batliner et al., 2010). Recent advances in language processing have provided automatic tools for phrase segmentation. This paper employs an implementation of the SVM based chunk identification algorithm proposed by Kudoh and Matsumoto (2000). Word level segmentation is also included in the analysis, even though its length may not be long enough to capture suprasegmental information. The word segmentation is obtained with a *hidden Markov model* (HMM) – based forced-alignment procedure (Lee et al., 2005).

A neutral corpus is used to build the neutral lexicon-independent functional PCA basis at sub-sentence levels (time-based, chunk or word). This corpus corresponds to the *Wall Street Journal-based Continuous Speech Recognition Corpus Phase II* (WSJ1) (Paul and Baker, 1992) (see Table 1). Only the spontaneous recordings corresponding to 8104 utterances pronounced by 50 subjects with varying degrees of experience in dictation were considered. First, the data is segmented according to each sub-sentence unit. Then, 200 utterances are randomly chosen across the 50 speakers to extract the corresponding segments (time-based windows, chunks or words). These 200 utterances generate over 1500 neutral segments for each of the segmentation units. The sentences are used to build the functional PCA bases.

Table 6
Accuracy with different segmentation units with lexicon-independent bases (EMA and EMO-DB). In EMO-DB database, Chunk and Word level are not provided since the phoneme boundary segmentation of speech is not available. We report the accuracies of the benchmark system trained with F0 statistics for sentence and time-based segmentations.

		Proposed system Segmentation level				Benchmark system Segmentation level			
	Emotion	Sentence	Time-based	Chunk	Word	Sentence	Time-based	Chunk	Word
EMA	Neutral-happiness	0.893	0.877	0.853	0.817	0.843	0.673	0.659	0.612
	Neutral-anger	0.795	0.790	0.790	0.733	0.737	0.632	0.619	0.592
	Neutral-sadness	0.547	0.553	0.480	0.470	0.653	0.581	0.560	0.544
	Neutral-emotional	0.742	0.744	0.773	0.745	0.714	0.668	0.667	0.629
EMO-DB	Neutral-fear	0.709	0.580	–	–	0.642	0.578	–	–
	Neutral-disgust	0.711	0.750	–	–	0.658	0.544	–	–
	Neutral-happiness	0.789	0.694	–	–	0.763	0.610	–	–
	Neutral-boredom	0.706	0.595	–	–	0.513	0.502	–	–
	Neutral-sadness	0.663	0.722	–	–	0.706	0.566	–	–
	Neutral-anger	0.777	0.706	–	–	0.825	0.615	–	–
	Neutral-emotional	0.713	0.744	–	–	0.690	0.523	–	–

Their F0 contours are linearly time-warped to the average segment duration. The resulting set of F0 contours was employed as input to estimate the functional PCA described in Fig. 2(a). This evaluation is conducted using the EMA and EMO-DB databases. Since we do not have word segmentation for the EMO-DB database, we only report results on the time-based segmentation.

The testing utterances are segmented according to the sub-sentence units. After extracting and post-processing their F0 contours, the projections onto the functional PCA basis are estimated for each segment (Fig. 2(b)). The QDC classifiers, which are trained at the sub-sentence levels, are employed to classify each testing segment using a leave-one-speaker-out scheme.

Table 6 presents the average performance for the time-based, chunk and word level segmentations. The results for sentence level with lexicon-independent references are also shown (values extracted from Table 5). Table 6 shows that in general the accuracies achieved at sentence level are higher than the ones achieved at sub-sentence levels. Time-based segmentation provides the highest emotional classification accuracy among the sub-sentence unit. As can be seen in Table 4, the differences in accuracy between sentence and time-based sub-sentence unit classifiers are not significant with all the emotional categories (EMA database). Similar results are observed with the EMO-DB database. However, more significant differences in accuracy are observed when sentence and word level based classifiers are compared. This result is consistent with our previous work which suggests that shorter speech units are not as effective to capture emotional information from the F0 contour (Busso et al., 2009). Table 6 provides the results for the benchmark system for sentence and time-based segmentations. The functional PCA projections consistently produces better performance than global F0 statistics. This result is particularly clear with time-based segmentation. Notice that as the window size is reduced, the statistics derived from the F0 contour are less reliable. In contrast, the functional PCA projections can capture local emotional variations conveyed in short segments.

5. Validation of the approach in a non-acted corpus

The proposed approach is validated with the spontaneous SEMAINE database (McKeown et al., 2010) (see Table 1 and Section 2.1). Instead of assigning an emotional label for each sentence, the subjective evaluations correspond to continuous assessment of the emotional content in real time using Feeltrace (50 values per second). Therefore, this database is ideal to evaluate whether the proposed approach can detect localized emotional information conveyed within the sentence. Previous studies have considered this database to recognize high and low values of valence, arousal, expectancy and power (Kim et al., 2011; Meng and Bianchi-Berthouze, 2011; Cen et al., 2011). They reported accuracies around 50% in binary classifications at word level. Since the evaluation setups were different, the reported performances cannot be directly compared with our results. However, these studies show the challenging task of recognizing the emotional state from this corpus.

Table 7
Inter-evaluator agreement and subjective-objective correlations with the SEMAINE database for different window length (IEA = correlation between the subjective evaluation of one subject and the mean of the remaining raters in the database, $\rho(S, O)$ = correlation between the average subjective evaluation and the objective metric, $\rho(\Delta S/\Delta t, O)$ = correlation between the derivative of the average subjective evaluation and the objective metric).

Window length (s)	IEA	$\rho(S, O)$	$\rho(\Delta S/\Delta t, O)$
0.25	0.347	0.235	0.327
0.50	0.351	0.250	0.396
1.00	0.363	0.215	0.438

We compare the similarity between an objective metric derived from the projections of the functional PCA basis with the average values of the subjective evaluations. The experiments were carried out by using time-based segmentation, as explained above. The WSJ1 database was employed to build the lexicon-independent neutral functional PCA basis. For each window, the projections into this basis are estimated. Motivated by the results in Fig. 4, we compute the norm of the projection, which is used as an objective metric of the emotional prominence conveyed in speech. This norm is smoothed with a median filter. As a ground truth, we estimated a subjective emotional prominence measure. While using Feeltrace, the evaluators are instructed to place the pointer in the center of the coordinate system to describe neutral speech. The distance of the pointer from the center is regarded as the emotional intensity of the speech (Cowie et al., 2000). Therefore, we define the subjective metric $e(t)$ as:

$$e(t) = \sqrt{a^2(t) + v^2(t)} \tag{7}$$

where $a(t)$ and $v(t)$ are the average activation and valence subjective curves given by the human evaluators.

The use of Feeltrace as an evaluation tool introduces some challenges. The raters have to sense the stimuli, perceive the message, decide its emotional attributes and move the pointer according to their perceptual judgments, all this in real time. This perceptual process introduces a delay that is intrinsically speaker independent (Eyben et al., 2010). However, the proposed approach captures the instantaneous emotional content in the signal. The lag between the signals is addressed by allowing a variable shift between the objective and subjective metrics. This lag is estimated for each sentence by maximizing the correlation between the metrics. The lag is forced to be lower than 0.5 s (a similar threshold is proposed in Nicolaou et al. (2011) to address the synchronization issue).

Table 7 presents the average Pearson's correlation between the subjective and objective metrics describing the emotional prominence in the SEMAINE database (column $\rho(S, O)$). These results are estimated with time-based segments of various lengths (0.25, 0.5 and 1 s), with 50% of overlap. The averaged correlation between the objective and subjective metrics is $\rho = 0.25$, when the segment window is set to 0.5 s. As a comparison, Table 7 also shows the *inter-evaluator agreement* (IEA) estimated as the averaged correlation between the curves of one rater and the averaged curves of the other raters (column *IEA*). The inter-evaluator agreement is $\rho = 0.35$ when the segment window is set to 0.5 s. Although our proposed method provides a lower correlation, the proposed objective metric approaches the correlation observed between evaluators. Notice that this comparison is not completely fair, since the evaluators made their judgments after watching the video and listening to the speech. In contrast, the proposed objective metric is estimated by making use of only F0 contours. Fig. 5(a) presents an example with subjective and objective metrics for a given utterance in the SEMAINE database (the x-axis corresponds to time). For graphical reasons, a constant normalization factor was applied to the objective measure. For this example, the correlation between both curves is $\rho = 0.51$. This figure also shows the lag between the metrics.

We noticed that in certain sentences, the correlation between the subjective and objective metrics was low or even negative. Fig. 5(b) shows an example in which the correlation between the curves was $\rho = -0.24$. An interesting pattern in this figure is the cumulative behavior of the subjective curve, which was also observed in other sentences. We hypothesize that, after perceiving a localized emotional segment with high intensity, human evaluators tend to keep the cursor position in the same place for some time even though the intrinsic emotional intensity decreases. This cumulative behavior also agree with studies that show that individuals are more sensitive to relative variations in the emotional intensity (Russell and Fehr, 1987). In fact, the highest variation in the subjective curve shown in Fig. 5(b) coincides with the highest emotional prominence measured by the proposed objective metric. By considering these findings, we decide to compare the proposed objective metric with the derivative of the subjective curves (i.e., variations rather than absolute values). For example, Fig. 5(b) shows the derivative of the subjective evaluation (dashed line). The correlation

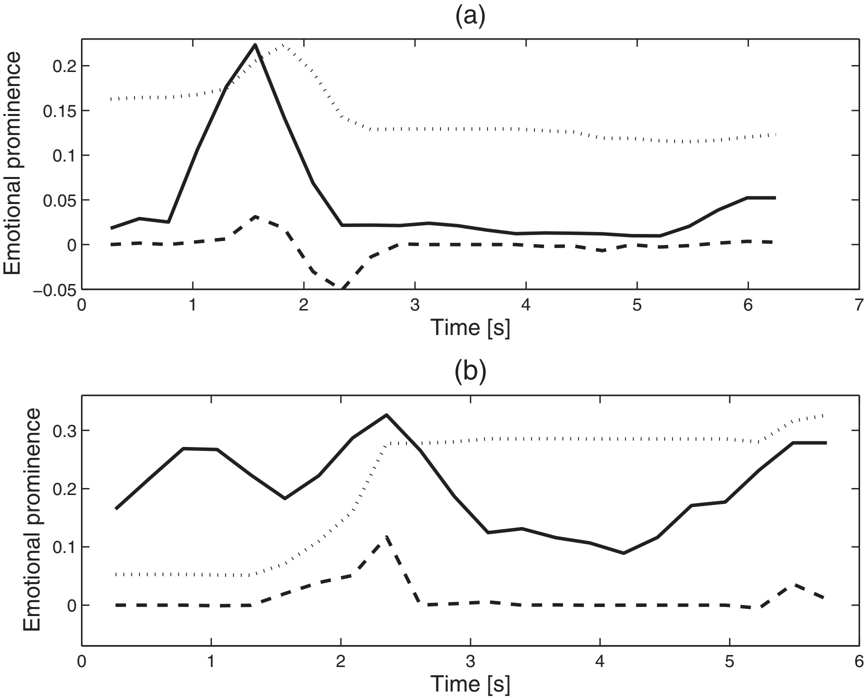

Fig. 5. (a) Example of subjective (dotted), derivative of subjective (dashed) and objective (solid) metrics. In this example, the correlation between subjective and objective metrics is equal to $\rho = 0.51$. (b) Example of subjective (dotted), derivative of subjective (dashed) and objective (solid) metrics. In this example, the correlation between the derivative of subjective and objective metrics is equal to $\rho = 0.55$.

between this signal and the proposed metric is $\rho = 0.55$. Table 7 shows the correlation between the proposed objective metric and the derivative of the subjective metrics across all the sentences (column $\rho(\Delta S/\Delta t, O)$). The correlations are higher than when absolute values of the subjective curves are employed (column $\rho(S, O)$). Interestingly, the correlation values are even higher than the inter-evaluation agreement, when the window lengths is set to 0.5 or 1 s.

Finally, we use the values of activation and valence given by the evaluators to create neutral and emotional classes. As mentioned, the center of the coordinate in Feeltrace corresponds to neutral speech. Therefore, if the subjective metric $e(t)$, defined in Eq. (7), is smaller than a threshold (e_{th}) the segment is considered as neutral. If $e(t)$ is outside this circle, the segment is considered emotional. We set $e_{th} = 0.3$, which produces balanced classes (neutral/emotional ratio equals to 1.1357). Notice that similar threshold was used by Schuller et al. (2011). With these classes, we conduct binary classification experiments using time-based segmentation. The window size of the segments is 0.5 s. The WSJ1 database is employed to build the lexicon-independent neutral functional PCA basis. We estimate the projections into this basis, which are used to classify the speech segments as neutral or emotional with a QDC classifier. Table 8 compares the results achieved by the proposed method and by the benchmark classifier trained with F0 statistics. The proposed approach achieves an accuracy of 62.7% which is 5% higher than the accuracy of the benchmark method. This result suggests that the proposed method captures emotional information conveyed on the F0 contour producing better performance than classifiers trained with features extracted from F0 statistics.

Table 8
Binary classification (SEMAINE). The classes are created by thresholding the activation/valence scores. We report the accuracy, average precision, average recall and *F*-score.

	Accuracy	Average precision	Average recall	*F*-score
Proposed system	0.621	0.618	0.623	0.618
Benchmark	0.571	0.574	0.575	0.564

6. Conclusions

This paper proposed a novel method to detect emotional modulation in F0 contours by using neutral reference models with functional PCA basis functions. The projections into this basis define the features that are used to train an emotion detection system. The approach was evaluated under different conditions. First, we built lexicon-dependent conditions (i.e., one basis per sentence), which achieved accuracies as high as 75.8% in binary emotion classification tasks. This performance is 6.2% higher than the ones achieved by a benchmark classifier trained with global F0 statistics. Then, we evaluated lexicon-independent functional PCA basis, built with F0 contours extracted from utterances with different lexical content. The results showed that the degradation in accuracy provided by lexicon-independent models was not significant when compared with the lexicon-dependent system (i.e. from 75.8% to 74.2%). The proposed system was then applied at the sub-sentence level to detect the most emotionally salient segments. The difference in accuracy between sentence and time-based segment classifiers was not significant across all the emotional categories. Finally, the approach was validated with a spontaneous database. The objective emotional prominence metric given by the functional PCA projections correlates with subjective evaluations. Furthermore, the system achieved 62.7% accuracy in binary classification which is 5% higher than benchmark classifier.

Our future work includes the incorporation of other prosodic features such as energy contour and duration. Likewise, the proposed scheme can be extended to detect specific emotional categories (e.g., happiness versus anger). For example, we can build emotion dependent functional PCA basis with F0 contours extracted from sentences labeled with a target emotion. Alternatively, the emotion detection system can be used as a first step in a more sophisticated multi-class emotion recognition system, in which emotional speech samples are further assigned to finer emotional labels (e.g., happiness versus anger). The combination of FDA with the polynomial representation of curves can also be proposed as future research. Finally, the functional PCA based method presented in this paper can be even extended to other speech processing tasks such as prosody assessment in second language learning. Basically, the idea is to train a functional PCA basis by using prosodic patterns extracted from native speech. Then, the testing F0 contours generated by non-native speakers can be assessed by employing the projections onto the native speaker basis.

Acknowledgment

This work was funded by the Government of Chile under grants Fondecyt 1100195 and Mecesup FSM0601, and US National Science Foundation under grants IIS-1217104 and IIS-1329659.

References

Arias, J., Yoma, N., Vivanco, H., 2010. Automatic intonation assessment for computer aided language learning. Speech Communication 52, 254–267.

Batliner, A., Seppi, D., Steidl, S., Schuller, B., 2010. Segmenting into adequate units for automatic recognition of emotion-related episodes: a speech-based approach. Advances in Human-Computer Interaction 2010, 1–15.

Boersma, P., Weenink, D., 1996. Praat, a system for doing phonetics by computer. Technical Report 132. Institute of Phonetic Sciences of the University of Amsterdam, Amsterdam, Netherlands http://www.praat.org

Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W., Weiss, B., 2005. A database of German emotional speech. In: 9th European Conference on Speech Communication and Technology (Interspeech'2005 - Eurospeech), Lisbon, Portugal, pp. 1517–1520.

Burleson, W., Picard, R., 2004. Affective agents: sustaining motivation to learn through failure and a state of "stuck". In: Social and Emotional Intelligence in Learning Environments Workshop in conjunction with the 7th International Conference on Intelligent Tutoring Systems (ITS 2004), Maceiò, Brazil.

Busso, C., Bulut, M., Narayanan, S., 2013. Toward effective automatic recognition systems of emotion in speech. In: Gratch, J., Marsella, S. (Eds.), Social Emotions in Nature and Artifact: Emotions in Human and Human–Computer Interaction. Oxford University Press, New York, NY, USA.

Busso, C., Lee, S., Narayanan, S., 2009. Analysis of emotionally salient aspects of fundamental frequency for emotion detection. IEEE Transactions on Audio. Speech and Language Processing 17, 582–596.

Busso, C., Narayanan, S., 2007. Joint analysis of the emotional fingerprint in the face and speech: a single subject study. In: International Workshop on Multimedia Signal Processing (MMSP 2007), Chania, Crete, Greece, pp. 43–47.

Cen, L., Yu, Z.L., Dong, M., 2011. Speech emotion recognition system based on L1 regularized linear regression and decision fusion. In: DMello, S., Graesser, A., Schuller, B., Martin, J.C. (Eds.), Affective Computing and Intelligent Interaction (ACII 2011). Springer, Berlin /Heidelberg, Memphis, TN, USA, pp. 332–340, volume 6975/2011 of Lecture Notes in Computer Science.

Cowie, R., Douglas-Cowie, E., Savvidou, S., McMahon, E., Sawey, M., Schröder, M., 2000. 'FEELTRACE': an instrument for recording perceived emotion in real time. In: ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion, ISCA, Newcastle, Northern Ireland, UK, pp. 19–24.

Cowie, R., Douglas-Cowie, E., Tsapatsoulis, N., Votsis, G., Kollias, S., Fellenz, W., Taylor, J., 2001. Emotion recognition in human-computer interaction. IEEE Signal Processing Magazine 18, 32–80.

El Ayadi, M., Kamel, M., Karray, F., 2011. Survey on speech emotion recognition: Features, classification schemes, and databases. Pattern Recognition 44, 572–587.

Eyben, F., Wöllmer, M., Schuller, B., 2009. openEAR-introducing the munich open-source emotion and affect recognition toolkit. In: International Conference on Affective Computing and Intelligent Interaction (ACII 2009), Amsterdam, The Netherlands, pp. 576–581.

Eyben, F., Wöllmer, M., Graves, A., Schuller, B., Douglas-Cowie, E., Cowie, R., 2010. On-line emotion recognition in a 3-D activation-valence-time continuum using acoustic and linguistic cues. Journal on Multimodal User Interfaces 3, 7–19.

Grimm, M., Kroschel, K., Mower, E., Narayanan, S., 2007. Primitives-based evaluation and estimation of emotions in speech. Speech Communication 49, 787–800.

Gubian, M., Torreira, F., Strik, H., Boves, L., 2009. Functional data analysis as a tool for analyzing speech dynamics. a case study on the french word c'était. In: Interspeech 2009, Brighton, UK, pp. 2199–2202.

Jeon, J., Xia, R., Liu, Y., 2011. Sentence level emotion recognition based on decisions from subsentence segments. In: International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2011), Prague, Czech Republic, pp. 4940–4943.

Kim, J.C., Rao, H., Clements, M., 2011. Investigating the use of formant based features for detection of affective dimensions in speech. In: DMello, S., Graesser, A., Schuller, B., Martin, J.C. (Eds.), Affective Computing and Intelligent Interaction (ACII 2011). Springer, Berlin /Heidelberg, Memphis, TN, USA, pp. 369–377, volume 6975/2011 of Lecture Notes in Computer Science.

Koolagudi, S., Rao, K.S., 2012. Emotion recognition from speech: a review. International Journal of Speech Technology 15, 99–117.

Kudoh, T., Matsumoto, Y., 2000. Use of support vector learning for chunk identification. In: Workshop on Learning language in logic (LLL 2000) and conference on Computational natural language learning (CoNLL 2000), Lisbon, Portugal.

Langenecker, S., Bieliauskas, L., Rapport, L., Zubieta, J., Wilde, E.S., Berent, S., 2005. Face emotion perception and executive functioning deficits in depression. Journal of Clinical and Experimental Neuropsychology 27, 320–333.

Lee, C., Yildirim, S., Bulut, M., Kazemzadeh, A., Busso, C., Deng, Z., Lee, S., Narayanan, S., 2004. Emotion recognition based on phoneme classes. In: 8th International Conference on Spoken Language Processing (ICSLP 04), Jeju Island, Korea, pp. 889–892.

Lee, S., Yildirim, S., Kazemzadeh, A., Narayanan, S., 2005. An articulatory study of emotional speech production. In: 9th European Conference on Speech Communication and Technology (Interspeech'2005 – Eurospeech), Lisbon, Portugal, pp. 497–500.

Lieberman, P., Michaels, S., 1962. Some aspects of fundamental frequency and envelope amplitude as related to the emotional content of speech. Journal of the Acoustical Society of America 34, 922–927.

Liscombe, J., Venditti, J., Hirschberg, J., 2003. Classifying subject ratings of emotional speech using acoustic features. In: 8th European Conference on Speech Communication and Technology (EUROSPEECH 2003), Geneva, Switzerland, pp. 725–728.

McKeown, G., Valstar, M., Cowie, R., Pantic, M., 2010. The SEMAINE corpus of emotionally coloured character interactions. In: IEEE International Conference on Multimedia and Expo (ICME 2010), Singapore, pp. 1079–1084.

Meng, H., Bianchi-Berthouze, N., 2011. Naturalistic affective expression classification by a multi-stage approach based on hidden Markov models. In: DMello, S., Graesser, A., Schuller, B., Martin, J.C. (Eds.), Affective Computing and Intelligent Interaction (ACII 2011). Springer, Berlin /Heidelberg, Memphis, TN, USA, pp. 378–387, volume 6975/2011 of Lecture Notes in Computer Science.

Nicolaou, M., Gunes, H., Pantic, M., 2011. Continuous prediction of spontaneous affect from multiple cues and modalities in valence-arousal space. IEEE Transactions on Affective Computing 2, 92–105.

Paeschke, A., 2004. Global trend of fundamental frequency in emotional speech. In: Speech Prosody (SP 2004), Nara, Japan, pp. 671–674.

Paeschke, A., Sendlmeier, W., 2000. Prosodic characteristics of emotional speech: Measurements of fundamental frequency movements. In: ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion, Newcastle, Northern Ireland, UK, pp. 75–80.

Patterson, D., Ladd, D., 1999. Pitch range modelling: Linguistic dimensions of variation. In: Proceedings of the XIVth International Congress of Phonetic Sciences (ICPhS 1999), San Francisco, CA, USA, pp. 1169–1172.

Paul, D., Baker, J., 1992. The design for the Wall Street Journal-based CSR corpus. In: 2th International Conference on Spoken Language Processing (ICSLP 1992), Banff, Alberta, Canada, pp. 899–902.

Picard, R., 1997. Affective Computing. MIT Press, Cambridge, MA, USA.

Ramsay, J.O., Silverman, B.W., 2005. Functional Data Analysis. Springer Verlag, New York, NY, USA.

Rotaru, M., Litman, D.J., 2005. Using word-level pitch features to better predict student emotions during spoken tutoring dialogues. In: 9th European Conference on Speech Communication and Technology (Interspeech'2005 – Eurospeech), Lisbon, Portugal, pp. 881–884.

Russell, J., Fehr, B., 1987. Relativity in the perception of emotion in facial expressions. Journal of Experimental Psych: General 116, 223–237.

Schuller, B., Batliner, A., Steidl, S., Seppi, D., 2011a. Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge. Speech Communication 53, 1062–1087.

Schuller, B., Steidl, S., Batliner, A., Burkhardt, F., Devillers, L., Muller, C., Narayanan, S., 2010. The INTERSPEECH 2010 paralinguistic challenge. In: Interspeech 2010, Makuhari, Japan, pp. 2794–2797.

Schuller, B., Valstar, M., Eyben, F., McKeown, G., Cowie, R., Pantic, M., 2011. AVEC 2011- the first international audio/visual emotion challenge. In: DMello, S., Graesser, A., Schuller, B., Martin, J.C. (Eds.), Affective Computing and Intelligent Interaction (ACII 2011). Springer Berlin, Heidelberg, Memphis, TN, USA, pp. 415–424, volume 6975/2011 of Lecture Notes in Computer Science.

Silverman, K., Beckman, M., Pitrelli, J., Ostendorf, M., Wightman, C., Price, P., Pierrehumbert, J., Hirschberg, J., 1992. ToBI: a standard for labelling english prosody. In: 2th International Conference on Spoken Language Processing(ICSLP 1992), Banff, Alberta, Canada, pp. 867–870.

Taylor, P., 2000. Analysis and synthesis of intonation using the tilt model. Journal of the Acoustical Society of America 107, 1697–1714.

Vlasenko, B., Schuller, B., Mengistu, K.T., Rigoll, G., Wendemuth, A., 2008. Balancing spoken content adaptation and unit length in the recognition of emotion and interest. In: Interspeech 2008, Brisbane, Australia, pp. 805–808.

Wang, H., Li, A., Fang, Q., 2005. F0 contour of prosodic word in happy speech of Mandarin. In: Tao, J., Tan, T., Picard, R. (Eds.), Affective Computing and Intelligent Interaction (ACII 2005), Lecture Notes in Artificial Intelligence 3784. Springer-Verlag Press, Berlin, Germany, pp. 433–440.

Yang, L., Campbell, N., 2001. Linking form to meaning: the expression and recognition of emotions through prosody. In: ISCA ITRW on Speech Synthesis, Perthshire, Scotland.

Yildirim, S., Bulut, M., Lee, C., Kazemzadeh, A., Busso, C., Deng, Z., Lee, S., Narayanan, S., 2004. An acoustic study of emotions expressed in speech. In: 8th International Conference on Spoken Language Processing (ICSLP 04), Jeju Island, Korea, pp. 2193–2196.

Zellers, M., Gubian, M., Post, B., 2010. Redescribing intonational categories with functional data analysis. In: Interspeech 2010, Makuhari, Japan, pp. 1141–1144.

Zeng, Z., Pantic, M., Roisman, G., Huang, T., 2009. A survey of affect recognition methods: audio, visual, and spontaneous expressions. IEEE Transactions on Pattern Analysis and Machine Intelligence 31, 39–58.