Energy and F0 contour modeling with Functional Data Analysis for Emotional Speech Detection

Juan Pablo Arias[1], *Carlos Busso*[2] *and Nestor Becerra Yoma* [1]

[1]Speech Processing and Transmission Laboratory, Dept. of Electrical Engineering, Universidad de Chile, Santiago, Chile

[2]Multimodal Signal Processing (MSP) Laboratory, Dept. of Electrical Engineering, The University of Texas at Dallas, Richardson, TX 75080, USA

juaarias@ing.uchile.cl, busso@utdallas.edu, nbecerra@ing.uchile.cl

Abstract

This paper proposes the use of reference models to detect emotional prominence in the energy and F0 contours. The proposed framework aims to model the intrinsic variability of these prosodic features. We present a novel approach based on *Functional Data Analysis* (FDA) to build reference models using a family of energy and F0 contours, which are implemented with lexicon-independent models. The neutral models are represented by bases of functions and the testing energy and F0 contours are characterized by their projections onto the corresponding bases. The proposed system can lead to accuracies as high as 80.4% in binary emotion classification in the EMO-DB corpus, which is 17.6% higher than the one achieved by a benchmark classifier trained with sentence level prosodic features. The approach is also evaluated with the SEMAINE corpus, showing that it can be effectively used in real applications.
Index Terms: Emotion detection, prosody modeling, emotional speech analysis, expressive speech, functional data analysis.

1. Introduction

Expressive communication is a key aspect of human interaction. *Human machine interfaces* (HMIs) able to recognize expressive behaviors have the potential to engage the user in a more effective manner. Among many features, speech prosody provides relevant information to characterize the externalization of emotions. Changes in intonation, loudness and timing are modulated to express emotion [1, 2, 3]. As a result, features extracted from pitch, energy and duration (i.e., the acoustic correlates of prosody) have been widely employed to study the emotional modulation in speech [1]. The state-of-the-art approach in recognizing and detecting emotions consists in computing a set of global statistics or functionals such as mean, variance, range, maximum and minimum, extracted from low level descriptors (e.g., F0 contour and energy). Then, feature selection algorithms are employed to choose a subset with the most emotionally salient parameters. One limitation of global statistics is that they do not capture the shape of F0/energy contours, which could provide useful information for emotion detection. For example, low variations in the fundamental frequency can be subjectively relevant in the identification of emotions [4].

This paper presents a novel approach to detect emotional prominence by modeling the shape of the energy and F0 contours. The method generates emotionally neutral reference models for these prosodic features, which are used to contrast the testing sentence. These models correspond to bases that are built with *functional data analysis* (FDA) using emotionally neutral utterances from several speakers. Then, the testing energy and F0 contours are projected onto the reference bases, and their projections are used as features to discriminate between neutral and emotional speech. The proposed method is evaluated with the EMO-DB corpus, achieving accuracies as high as 80.4% in binary emotion classification tasks (i.e. neutral versus emotional speech), which is 17.6% higher than the accuracy achieved by a standard technique trained with global statistics from the energy and F0 contours. The proposed method is also evaluated with the spontaneous SEMAINE database, achieving an accuracy that is 7.7% better than the one achieved by the benchmark system. The proposed functional PCA projection captures deviations from neutral speech even when small segments are used (e.g., 0.5 sec windows).

2. Background

2.1. Related Work

Previous studies have attempted to model the shape of the F0 contour in the context of emotion. Paeschke and Sendlmeier [5] analyzed the rising and falling movements of the F0 contour 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 emotion classes. Rotaru and Litman employed linear and quadratic regression coefficients and regression error as features to represent pitch curves [6]. Yang and Campbell argued that concavity and convexity of the F0 contour reflect the underlying expressive state [7]. Liscombe et al. [8] used pitch accents and boundary tones labels from the *Tone and Break Indices system* (ToBI) [9] to recognize emotions.

Building upon our previous work on detecting expressive speech using neutral reference models [1, 10], this paper presents a novel framework to characterize the temporal shape of prosodic features. We create FDA-based reference models, which are used to contrast emotional speech. The approach is radically different from current approaches to recognize emotions, and provides an elegant solution to capture the local variability in the prosodic contours caused by expressive speech.

2.2. Emotional Database and Feature Extraction

The proposed method is evaluated with two publicly available emotional corpora that have been widely used in related work. Therefore, other groups can reproduce our results. We consider the acted corpus *Berlin Database of Emotional Speech* (EMO-DB) [11]. The database consists of ten speaker (five male and five female), who read ten German sentences one time expressing fear, disgust, happiness, boredom, sadness, and anger, in addition to neutral state. The second corpus is the SEMAINE database, which is a spontaneous non-acted emotional corpus (details are given in [12]). The emotions are annotated in terms of continuous emotional attributes, from which we con-

10.21437/Interspeech.2013-253

sider the activation (calm versus active) and valence (negative versus positive) dimensions. External evaluators assessed the emotional content of the corpus using the Feeltrace toolkit [13], which gives continuously over time values (50 values per second). The approach differs from conventional schemes of assigning one label per sentence. We use data from 10 speakers.

The fundamental frequency is estimated using the autocorrelation pitch detector implemented in Praat [14] (25ms frames with 50% overlap). We represent the F0 contour using a semitone scale and unvoiced segments are interpolated with cubic spline to obtain smooth and continuous trajectories. The resulting interpolated curve is normalized by subtracting the mean. Henceforth, the term "F0 contour" denotes the F0 curve in the semitone scale after interpolation and mean normalization. Likewise, the RMS energy contour, $E(t)$, is estimated for each frame. The resulting curve is represented in terms of decibels and normalized with respect to the mean energy. As a final step, the energy values lower than -15 dB are set to zero, since they introduces perturbation in the proposed models (Sec. 3).

3. Shape modeling approach with FDA

3.1. Functional Data Analysis (FDA)

FDA represents the structure of signals as functions, instead of data points [15]. The time series data is modeled as a continuous, smooth function, $x(t)$, created as a linear combination of basis functions ϕ_k:

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

where K is the dimension of the expansion and c_k is the projection onto the k-th basis function. Both ϕ_k and K are parameters of FDA that should be properly chosen according to the characteristics of the data. Functional data is observed as a discrete sequence (t_j, y_j), $j \in \{1, \ldots, n\}$, where y_j is the sampled value of the function $x(t)$ at time t_j. This sequence is not necessarily equally-distributed and may be corrupted by noise ϵ_j, i.e. $y_j = x(t_j) + \epsilon_j$. The process of fitting functions to data is known as *smoothing*. Given the discrete observations y_j and the basis functions $\{\phi_1, \ldots, \phi_K\}$, smoothing attempts to find coefficients c_k by minimizing the mean squared error ϵ_j. The optimal coefficients, $\hat{c_k}$, are estimated with Eq. 2,

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

where λ is a smoothing parameter and D^m represents the m-th derivative [15]. FDA provides several advantages when compared with classical approaches that represent data as a set of discrete samples. For example, powerful tools for analyzing data such as *principal component analysis* (PCA) can be used in the framework of FDA. Functional PCA extends the conventional PCA framework to the functions' domain [15]. Given a set of functions, denoted by $x_v(t)$, the principal components projections, $f_{u,v}$, are given by

$$f_{u,v} = \int \xi_u(t) x_v(t) dt \quad (3)$$

where $\xi_u(t)$ is an orthonormal basis denoted as *principal component* (PC) functions that represents the variability of $x_v(t)$. The function $x_v(t)$ can be approximated with the first U PCs:

$$\hat{x}_v(t) = \sum_{u=1}^{U} f_{u,v} \xi_u(t). \quad (4)$$

Functional PCA allows us to statistically represent a family of functions, which can be employed as neutral reference to contrast emotional speech. Previous studies have used FDA to provide a descriptive analysis of prosodic features [16, 17]. In contrast, this paper proposes FDA as a tool for generating neutral reference models to capture deviations from normal speech.

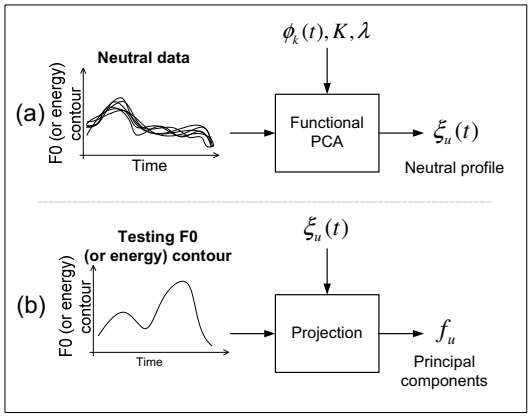


Figure 1: General framework: (a) neutral model with functional PCA; (b) projection of speech onto the neutral eigenspace.

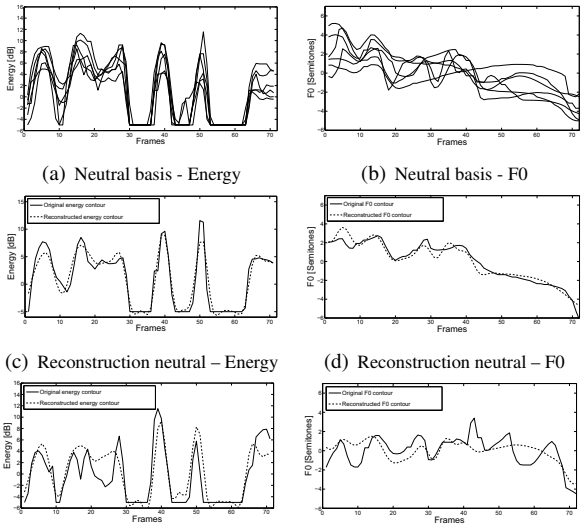
3.2. Proposed Approach

Figure 1-(a) describes the general framework to build the neutral reference model by employing 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 *dynamic time warping* (DTW). Then, the energy and F0 contours extraction procedure described in section 2.2 is applied to the signals. The resulting time-aligned, post-processed energy and F0 contours are smoothed and represented as functional data by employing a basis of B-spline functions $\phi_k(t)$ according to Eqs. (1) and (2). Finally, functional PCA is applied to generate a new orthogonal basis of functions $\xi_u(t)$.

Figure 1-(b) shows the testing stage of the proposed scheme. First, the testing speech is aligned with the training data using DTW. Then, we extract the energy and F0 contours which are projected onto the neutral reference bases $\xi_u(t)$. As a result, the coefficients f_u are obtained, which correspond to the parameters that describe the shape of the testing energy and F0 contours. Since the profile $\xi_u(t)$ is generated with non-emotional speech, it is expected that neutral and emotional testing energy and F0 contours will provide different projections (i.e. $\{f_1 \ldots f_U\}$) onto the functional PCA basis. Therefore, we propose to use the parameters $\{f_1 \ldots f_U\}$ as features to detect emotional speech.

Figure 2 presents an example of the approach for one of the ten sentences of the EMO-DB corpus. Figures 2(a) and 2(b) show the time-aligned and post-processed energy and F0 curves of six neutral realizations from different speakers. Although the sentences present variations in their prosodic contours, they clearly have a pattern that our approach aims to capture. A neutral profile is trained for this data with the proposed approach. The basis ϕ_k is implemented with a 6th order B-spline with $K = 40$. Figures 2(c) and 2(d), and figures 2(e) and 2(f) show the reconstruction of neutral and happy prosody contours, respectively, for the same sentence uttered by another subject (not considered for building the neutral reference). The prosodic curves are reconstructed using the first five PCs ($U = 5$). The figures show that the neutral F0 and energy contours are accurately approximated with the neutral functional PCA basis. The corresponding contours for happy speech are less accurate.

Figure 3 shows the average absolute value of the projec-



(a) Neutral basis - Energy (b) Neutral basis - F0

(c) Reconstruction neutral – Energy (d) Reconstruction neutral – F0

(e) Reconstruction happiness – Energy (f) Reconstruction happiness – F0

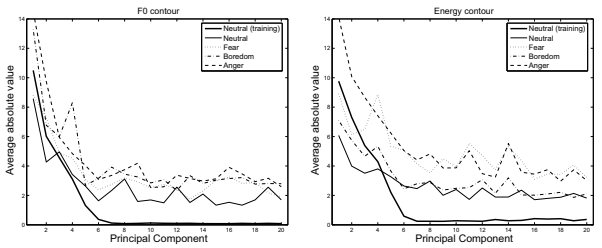
Figure 2: Reconstruction of energy and F0 contours with functional PCA: (a)-(b) training data to generate the neutral bases; (c)-(d) reconstruction of a neutral speech with five PCs; (e)-(f) reconstruction of a happy speech with five PCs.

tions onto the first 20 PCs for neutral and emotional speech (fear, boredom and anger). The functional PCA models are trained and tested with data from six and four speakers, respectively (speaker independent partitions). The figure shows that the mean of the contours' projections for the training neutral speech is approximately equal to zero when $k \geq 7$. Figure 3 also shows that when the energy contour is employed, the mean of the absolute value of the projections for fear and angry speech are higher than the neutral ones, even for higher order principal components. For $6 \leq k \leq 40$, the values of the projections for neutral speech converges to zero faster than the ones for emotional speech. The reference model fits better the prosodic features of the neutral sentences than the ones for emotional sentences. The differences in the projections estimated from prosody contours for neutral and emotional speech indicate that they can be considered as features to detect emotional speech.

4. Discriminant Analysis

To assess the discriminative power of the functional PCA projections, the proposed system is employed to detect emotional from neutral speech (i.e., binary problems). This approach is more general than classifying specific emotional classes, since it less dependent on the specific emotional classes of a corpus. It can also be used as the first step in a multi-class problem, in which a second classifier is trained to recognize the particular label for emotional samples detected by our system. We implement the classifiers with *quadratic discriminant classifier* (QDC). In our preliminary analysis, we considered *support support vector machine* (SVM), which requires a development set to identify the optimal kernel and soft margin parameter. In contrast, QDC does not require a validation partition to set the parameters and its performance is similar to SVM for this task.

We evaluate the approach with the EMO-DB (Sec. 4.1) and SEMAINE (Sec. 4.2) corpora. The example presented in section 3.2 uses one set of functional PCA bases per each of the 10 sentences in the EMO-DB corpus (i.e., lexicon-dependent (LD) bases). This approach is not practical for real applications. Therefore, the evaluation of the system is implemented with lexicon-independent (LI) bases, built with neutral sentences conveying different lexical content. Figures 2(a) and 2(b) suggest that lexical information affects the energy and F0 contours. 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. Therefore, robust reference models can be built. The energy and F0 contours are extracted and post processed from neutral sentences. Then, the average duration of the signals is estimated and used to linearly warp their energy and F0 contours which are employed as input to estimate functional PCA (Fig. 1-(a)). Then, the functional PCA projections are estimated and used as features (Fig. 1-(b)).



(a) F0 contour (b) Energy contour

Figure 3: Average absolute value of the projections associated to each principal component obtained with EMO-DB database.

4.1. Evaluation with EMO-DB Database

The EMO-DB database is 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 different speakers to ensure that the results are speaker independent. Only neutral data is employed to build the reference models. To maximize the use of the EMO-DB database, six permutations are implemented by interchanging the role of each partition among development, training and testing data sets. The performance rates are estimated by averaging the results obtained in all six implementations. The following binary classifiers are individually trained: neutral-fear, neutral-disgust, neutral-happiness, neutral-boredom, neutral-sadness and neutral-anger. In addition, we formed the "emotional" class by grouping utterances from the six emotional classes. A subset of utterances from each emotional class is randomly chosen to match the number of neutral samples (chance = 50%). This procedure is repeated 100 times and the performance rates are averaged. The system is implemented considering the projections into the bases of three different combinations of prosodic features: F0 contour only; energy contour only; and, both F0 and energy together.

Table 1-(a) shows the performance of the proposed system with lexicon-independent models. The accuracies for neutral-happy and neutral-anger classification tasks are higher than 77% with the F0 contour only. When we consider only the energy contour, we achieve an accuracy of 84.2% for neutral-fear and neutral-disgust task. When F0 and energy features are combined, the accuracy of the neutral-emotional classifier is 80.4%, which is higher than the ones obtained with F0 or energy contours, separately. These results validate the proposed scheme.

For comparison purpose, a benchmark system is implemented for binary emotion detection that uses as features statistics from the energy and F0 contours. First, we estimate sentence-level functionals derived from prosody. The features corresponds to the subset of statistics from energy and F0 contours employed in the Interspeech 2010 Paralinguistic Challenge [18] (80 F0 features and 42 energy features). Then, *for-*

Table 1: Results with EMO-DB: (a) PCA projections with lexicon-independent bases (LI-FDA); (b) benchmark system. Chance is always 50%. *Acc* = Accuracy, *Pre* = Precision, *Rec* = Recall, and *F* = F-score (standard deviation for accuracy is given in brackets).

		F0 CONTOUR ONLY				ENERGY CONTOUR ONLY				F0 AND ENERGY			
		Acc [%]	Pre [%]	Rec [%]	F [%]	Acc [%]	Pre [%]	Rec [%]	F [%]	Acc [%]	Pre[%]	Rec [%]	F [%]
(a) LI-FDA	Neutral-Fear	70.9 (5.0)	73.7	67.3	68.5	84.2 (4.5)	87.3	80.2	83.4	83.9 (5.8)	8.66	81.3	82.7
	Neutral-Disgust	71.1 (4.9)	73.8	67.5	68.6	84.2 (4.5)	87.3	80.2	83.4	84.0 (5.7)	8.67	81.3	82.8
	Neutral-Happiness	78.9 (3.2)	78.8	80.6	79.3	81.8 (2.7)	87.0	75.9	80.4	88.6 (2.6)	90.9	86.6	88.3
	Neutral-Boredom	70.6 (5.7)	75.3	61.4	67.4	68.6 (3.7)	70.1	65.4	67.0	74.5 (4.7)	81.9	63.4	70.7
	Neutral-Sadness	66.3 (5.6)	80.7	43.2	55.7	68.6 (3.9)	69.6	66.4	67.4	70.5 (7.3)	90.3	46.2	59.6
	Neutral-Anger	77.7 (3.4)	78.0	78.5	77.9	95.0 (1.3)	96.7	93.2	94.9	92.9 (2.2)	94.5	91.3	92.8
	Neutral-Emotional	**71.3** (3.6)	**75.6**	**64.1**	**69.1**	**75.9** (1.6)	**80.0**	**69.2**	**74.2**	**80.4** (1.8)	**88.3**	**70.3**	**78.2**
(b) BENCHMARK	Neutral-Fear	64.2 (12.1)	86.1	37.0	46.9	70.7 (6.0)	71.9	69.1	70.2	60.1 (8.7)	95.5	21.6	44.8
	Neutral-Disgust	65.8 (10.8)	72.6	56.2	58.3	69.1 (6.2)	71.3	67.8	67.6	65.2 (12.7)	82.2	46.0	57.1
	Neutral-Happiness	76.3 (9.4)	96.4	55.2	67.9	69.1 (7.7)	68.3	75.0	70.9	79.2 (10.8)	97.2	60.4	71.9
	Neutral-Boredom	52.1 (1.8)	82.1	44.2	51.4	61.4 (3.9)	63.3	55.1	58.7	54.2 (8.4)	90.4	10.6	23.4
	Neutral-Sadness	70.6 (8.7)	64.4	96.6	76.9	77.1 (4.7)	75.4	80.9	77.9	73.1 (8.5)	66.2	98.3	78.8
	Neutral-Anger	82.5 (12.0)	93.0	70.9	77.4	87.2 (5.3)	87.7	87.0	87.1	85.3 (13.0)	94.8	72.4	79.9
	Neutral-Emotional	**69.0** (9.7)	**88.9**	**45.8**	**55.5**	**65.9** (7.3)	**67.3**	**67.5**	**66.6**	**62.8** (9.1)	**95.9**	**27.2**	**39.0**

ward feature selection (FFS) is applied to reduce the number of features to 20 (when we consider energy or F0 contour only), or 40 (with both F0 and energy together). These values are selected to match the number of projections used as features in the proposed approach. The database is divided into training (six speakers) and testing (four speakers). The experiments are carried out by employing the leave-four-speakers-out strategy. Table 1-(b) shows the benchmark results for the EMO-DB corpus. When compared with the benchmark system (Tab. 1-(b)), the LI approach leads to improvements in accuracy for the neutral-emotional task equal to 2.3%, 10.0% and 17.6% with F0 only, energy only and F0 plus energy, respectively (significant with *p-value*=0.272, *p-value*=0.040 and *p-value*$<$0.001, respectively). Similar improvements in performance are observed for most of the emotional classes. Furthermore, the standard deviations of the accuracy (numbers between brackets) obtained by the proposed method are much lower than the ones achieved with the benchmark system. This result suggests that the proposed technique is more consistent.

4.2. Evaluation with SEMAINE Database

We evaluate the proposed approach using the SEMAINE database. We conduct experiments with time-based segmentation, in which the corpus is split into windows of fix length (e.g., 1 sec segments). This approach is appealing for real-time applications, since the speech signal does not need to be pre-segmented. The corpus is evaluated with Feeltrace, which provides continuously over time values in the range -1 to 1. Similar to other studies, we transform the problem into binary classification by setting thresholds over the average scores across time and evaluators [19]. Given that the evaluators are instructed to provide scores close to zero for neutral speech, we define a circle centered in the origin of the activation-valence space (we used similar approach in [20]). The ratio of the circle is set to 0.3 which gives fairly balanced classes (neutral versus emotional speech). Studies addressing the problem of recognizing high and low values for each emotional dimension have shown accuracy around 53% [21, 22]. The low performance shows the challenges of working with this spontaneous corpus.

Since this is an spontaneous corpus, we can train the functional PCA bases using the spontaneous portion of the emotionally neutral WSJ1 corpus [23] (200 sentences randomly selected from 50 subjects). By training the bases with a separate neutral corpus, we do not need a development set. Therefore, the SEMAINE corpus is only used for training (5 speakers) and testing (5 speakers) the emotion detection classifiers. We use a speaker independent two-cross validation scheme, and we report the average results. We train the proposed classifier with the functional PCA projections, and a benchmark system with features derived from global statistics (energy and F0 contours).

Table 2: Results with SEMAINE (0.5 sec and 1 sec windows). *Acc* = Accuracy, *Pre* = Average precision, *Rec* = Average recall, and *F* = Average F-score.

	0.5 sec windows				1 sec windows			
	Acc	Pre	Rec	F	Acc	Pre	Rec	F
FDA (F0)	62.1	61.8	62.3	61.8	63.6	63.6	63.6	63.6
FDA (E)	56.7	57.2	57.0	56.4	57.6	57.1	59.0	57.0
FDA (E+F0)	63.1	63.7	63.7	63.1	64.2	64.3	64.2	64.2
Ben. (F0)	57.1	57.4	57.5	56.4	58.4	57.8	57.7	57.7
Ben. (E)	54.7	55.1	55.2	54.6	56.3	54.9	54.8	54.8
Ben. (E+F0)	55.4	55.5	55.6	55.0	57.4	56.5	56.3	56.3

Table 2 shows the results for both classifiers trained with 0.5 sec and 1 sec windows. In both cases, the classifiers trained with functional PCA show better performance than the ones trained with global statistics. When energy and F0 features are used, the improvement in accuracy for the proposed system is 7.7% (0.5 sec) and 6.9% (1 sec) over the benchmark system. While the accuracies of the proposed (-1.1%) and benchmark (-1.9%) systems drop when the 0.5 sec windows are considered, the classifiers trained with global statistics is the most affected. As the size of the window decreases, the estimation of global statistics is less robust, decreasing the consistency of the features. In contrast, the functional PCA projections capture deviation from neutral speech even in small segments.

5. CONCLUSIONS

This paper describes a novel method to detect emotional modulation in the energy and F0 contours by using neutral reference models defined as functional PCA bases. The proposed approach achieves better accuracies than a conventional approach trained with global statistics derived from the prosodic features for both the EMO-DB (17.6%) and SEMAINE (7.7%) corpora.

The future directions of this work include the evaluation of the approach with an extensive set of prosodic and spectral features. The FDA bases will be trained with a large neutral corpus producing robust reference models. We will use this approach to locally detect the most emotionally salient segments within a given utterance by employing shorter analysis windows. Finally, the functional PCA based method will be adapted to solve other speech processing tasks such as prosody assessment in second language learning [24] (capture the deviation of the user's intonation from the canonical pronunciation, described by the FDA model).

6. Acknowledgements

Work funded by the Government of Chile (Fondecyt 1100195, Mecesup FSM0601), and NSF (IIS-1217104, IIS-1329659).

7. References

[1] C. Busso, S. Lee, and S.S. Narayanan, "Analysis of emotionally salient aspects of fundamental frequency for emotion detection," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 17, no. 4, pp. 582–596, May 2009.

[2] T. Bänziger and K.R. Scherer, "The role of intonation in emotional expressions," *Speech Communication*, vol. 46, no. 3-4, pp. 252–267, July 2005.

[3] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J.G. Taylor, "Emotion recognition in human-computer interaction," *IEEE Signal Processing Magazine*, vol. 18, no. 1, pp. 32–80, January 2001.

[4] P. Lieberman and S.B. Michaels, "Some aspects of fundamental frequency and envelope amplitude as related to the emotional content of speech," *Journal of the Acoustical Society of America*, vol. 34, no. 7, pp. 922–927, July 1962.

[5] A. Paeschke and W.F. Sendlmeier, "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, September 2000, pp. 75–80.

[6] M. Rotaru and D. J. Litman, "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, September 2005, pp. 881–884.

[7] L. Yang and N. Campbell, "Linking form to meaning: the expression and recognition of emotions through prosody," in *ISCA ITRW on Speech Synthesis*, Perthshire, Scotland, August-September 2001.

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

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

[10] C. Busso, S. Lee, and S.S. Narayanan, "Using neutral speech models for emotional speech analysis," in *Interspeech 2007 - Eurospeech*, Antwerp, Belgium, August 2007, pp. 2225–2228.

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

[12] G. McKeown, M.F. Valstar, R. Cowie, and M. Pantic, "The SEMAINE corpus of emotionally coloured character interactions," in *IEEE International Conference on Multimedia and Expo (ICME 2010)*, Singapore, July 2010, pp. 1079–1084.

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

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

[15] J. O. Ramsay and B. W. Silverman, *Functional Data Analysis*, Springer Verlag, New York, NY, USA, 2005.

[16] M. Gubian, F. Torreira, H. Strik, and L. Boves, "Functional data analysis as a tool for analyzing speech dynamics. a case study on the french word c'était," in *Interspeech 2009*, Brighton, UK, September 2009, pp. 2199–2202.

[17] M. Zellers, M. Gubian, and B. Post, "Redescribing intonational categories with functional data analysis," in *Interspeech 2010*, Makuhari, Japan, September 2010, pp. 1141–1144.

[18] B. Schuller, S. Steidl, A. Batliner, F. Burkhardt, L. Devillers, C. Muller, and S. Narayanan, "The INTERSPEECH 2010 paralinguistic challenge," in *Interspeech 2010*, Makuhari, Japan, September 2010, pp. 2794–2797.

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

[20] T. Rahman and C. Busso, "A personalized emotion recognition system using an unsupervised feature adaptation scheme," in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2012)*, Kyoto, Japan, March 2012, pp. 5117–5120.

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

[22] M. Glodek, S. Tschechne, G. Layher, M. Schels, T. Brosch, S. Scherer, M. Kächele, M. Schmidt, H. Neumann, G. Palm, and F. Schwenker, "Multiple classifier systems for the classification of audio-visual emotional states," in *Affective Computing and Intelligent Interaction (ACII 2011)*, S. DMello, A. Graesser, B. Schuller, and J.-C. Martin, Eds., vol. 6975/2011 of *Lecture Notes in Computer Science*, pp. 359–368. Springer Berlin / Heidelberg, Memphis, TN, USA, October 2011.

[23] D.B. Paul and J.M. Baker, "The design for the Wall Street Journal-based CSR corpus," in *2th International Conference on Spoken Language Processing (ICSLP 1992)*, Banff, Alberta, Canada, October 1992, pp. 899–902.

[24] J.P. Arias, N.B. Yoma, and H. Vivanco, "Automatic intonation assessment for computer aided language learning," *Speech Communication*, vol. 52, no. 1, pp. 254–267, March 2010.