Analysis of Emotionally Salient Aspects of Fundamental Frequency for Emotion Detection

Carlos Busso, *Member, IEEE*, Sungbok Lee, *Member, IEEE*, and Shrikanth Narayanan, *Fellow, IEEE*

***Abstract*—During expressive speech, the voice is enriched to convey not only the intended semantic message but also the emotional state of the speaker. The pitch contour is one of the important properties of speech that is affected by this emotional modulation. Although pitch features have been commonly used to recognize emotions, it is not clear what aspects of the pitch contour are the most emotionally salient. This paper presents an analysis of the statistics derived from the pitch contour. First, pitch features derived from emotional speech samples are compared with the ones derived from neutral speech, by using symmetric Kullback–Leibler distance. Then, the emotionally discriminative power of the pitch features is quantified by comparing nested logistic regression models. The results indicate that gross pitch contour statistics such as mean, maximum, minimum, and range are more emotionally prominent than features describing the pitch shape. Also, analyzing the pitch statistics at the utterance level is found to be more accurate and robust than analyzing the pitch statistics for shorter speech regions (e.g., voiced segments). Finally, the best features are selected to build a binary emotion detection system for distinguishing between emotional versus neutral speech. A new two-step approach is proposed. In the first step, reference models for the pitch features are trained with neutral speech, and the input features are contrasted with the neutral model. In the second step, a fitness measure is used to assess whether the input speech is similar to, in the case of neutral speech, or different from, in the case of emotional speech, the reference models. The proposed approach is tested with four acted emotional databases spanning different emotional categories, recording settings, speakers and languages. The results show that the recognition accuracy of the system is over 77% just with the pitch features (baseline 50%). When compared to conventional classification schemes, the proposed approach performs better in terms of both accuracy and robustness.**

***Index Terms*—Emotional speech analysis, emotional speech recognition, expressive speech, intonation, pitch contour analysis.**

I. Introduction

EMOTION plays a crucial role in day-to-day interpersonal human interactions. Recent findings have suggested that emotion is integral to our rational and intelligent decisions. It helps us to relate with each other by expressing our feelings and providing feedback. This important aspect of human interaction needs to be considered in the design of *human–machine interfaces* (HMIs) [1]. To build interfaces that are more in tune with the users' needs and preferences, it is essential to study how emotion modulates and enhances the verbal and nonverbal channels in human communication.

Speech prosody is one of the important communicative channels that is influenced by and enriched with emotional modulation. The intonation, tone, timing, and energy of speech are all jointly influenced in a nontrivial manner to express the emotional message [2]. The standard approach in current emotion recognition systems is to compute high-level statistical information from prosodic features at the sentence-level such as mean, range, variance, maximum, and minimum of F0 and energy. These statistics are concatenated to create an aggregated feature vector. Then, a suitable feature selection technique, such as forward or backward feature selection, sequential forward floating search, genetic algorithms, evolutionary algorithms, linear discriminant analysis, or principal component analysis [3]–[5], is used to extract a feature subset that provides better discrimination for the given task. As a result, the selected features are sensitive to the training and testing conditions (database, emotional descriptors, recording environment). Therefore, it is not surprising that the models do not generalize across domains, and notably in real-life scenarios. A detailed study of the emotional modulation in these features can inform the development of robust features, not only for emotion recognition but also for other applications, such as expressive speech synthesis. This paper focuses on one aspect of expressive speech prosody: the F0 (pitch) contour.

The goal of this paper is twofold. The first is to study which aspects of the pitch contour are manipulated during expressive speech (e.g., curvature, contour, shape, dynamics). For this purpose, we present a novel framework based on *Kullback–Leibler divergence* (KLD) and logistic regression models to identify, quantify, and rank the most emotionally salient aspects of the F0 contour. Different acted emotional databases are used for the study, spanning different speakers, emotional categories and languages (English and German). First, the symmetric Kullback–Leibler distance is used to compare the distributions of different pitch statistics (e.g., mean, maximum) between emotional speech and reference neutral speech. Then, a logistic regression analysis is implemented to discriminate emotional speech from neutral speech using the pitch statistics as input. These experiments provide insights about the aspects of pitch that are modulated to convey emotional goals. The second goal is to use these emotionally salient features to build robust

Manuscript received June 22, 2008; revised October 31, 2008. Current version published March 10, 2009. This work was supported in part by the National Science Foundation (NSF) through the Integrated Media Systems Center, an NSF Engineering Research Center, under Cooperative Agreement EEC-9529152 and a CAREER award, in part by the Department of the Army, in part by the Office of Naval Research under a MURI award. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Mark Hasegawa-Johnson.

The authors are with the Signal and Image Processing Institute, University of Southern California, Los Angeles, CA 90089-2564 USA (e-mail: busso@usc.edu).

Digital Object Identifier 10.1109/TASL.2008.2009578

prosody speech models to detect emotional speech. In our recent work, we introduced the idea of building neutral speech models to discriminate emotional speech from neutral speech [6]. This approach is appealing since many neutral speech corpora are available, compared to emotional speech corpora, allowing the construction of robust neutral speech models. Furthermore, since these models are independent of the specific emotional databases, they can be more easily generalized to real-life applications [7]. While the focus on our previous paper was on spectral speech models, this paper focuses on features derived from the F0 contour. *Gaussian mixture models* (GMMs) are trained using the most discriminative aspects of the pitch contour, following the analysis results presented in this paper.

The results reveal that features that describe the global aspects (or properties) of the pitch contour, such as the mean, maximum, minimum, and range, are more emotionally salient than features that describe the pitch shape itself (e.g., slope, curvature, and inflexion). However, features such as pitch curvature provide complementary information that is useful for emotion discrimination. The classification results also indicate that the models trained with the statistics derived over the entire sentence have better performance in terms of accuracy and robustness than when they are trained with features estimated over shorter speech regions (e.g., voiced segments).

Using the most salient pitch features, the performance of the proposed approach for binary emotion recognition reaches over 77% (baseline 50%), when the various acted emotional databases are considered together. Furthermore, when the system is trained and tested with different databases (in a different language), the recognition accuracy does not decrease compared to the case without any mismatch between the training and testing condition. In contrast for the same task, the performance of a conventional emotion recognition system (without the neutral models) decreases up to 17.9% (absolute) using the same pitch features. These results indicate that the proposed GMM-based *neutral* model approach for binary emotion discrimination (emotional versus neutral speech) outperforms conventional emotion recognition schemes in terms of accuracy and robustness.

The paper is organized as follows. Section II provides the background and related work. Section III gives an overview of the proposed approach. It also describes the databases and the pitch features included in the analysis. Section IV presents the experiments and results based on KLD. Section V gives the experiments and results of the logistic regression analysis. Based on the results derived from previous sections of the paper, Section VI discusses the aspects of the pitch contour during expressive speech that are most distinctive, and therefore, most useful for emotion discrimination. Section VII presents the idea and the classification results of neutral reference models for expressive versus non-expressive speech classification. Finally, Section VIII gives the concluding remarks and our future research directions.

II. Related Work

Pitch features from expressive speech have been extensively analyzed during the last few years. From these studies, it is well known that the pitch contour presents distinctive patterns for certain emotional categories. In an exhaustive review, Juslin and Laukka reported some consistent results for the pitch contour across 104 studies on vocal expression [8]. For example, they concluded that the pitch contour is higher and more variable for emotions such as anger and happiness and lower and less variable for emotions such as sadness. Despite having a powerful descriptive value, these observations are not adequate to quantify the discriminative power and the variability of the pitch features. In this section, we highlight some of the studies that have attempted to measure the emotional information conveyed in different aspects of the pitch contour.

The results obtained by Lieberman and Michaels indicate that the fine structure of the pitch contour is an important emotional cue [9]. Using human perceptual experiments, they showed that the recognition of emotional modes such as bored and pompous decreased when the pitch contour is smoothed. Therefore, they concluded that small pitch fluctuations, which are usually neglected, convey emotional information.

In many languages, the F0 values tend to gradually decrease toward the end of the sentence, a phenomenon known as declination. Wang *et al.* compared the pitch declination conveyed in happy and neutral speech in Mandarin [10]. Using four-word sentences, they studied the pitch patterns at the word level. They concluded that the declination in happy speech is less than in neutral speech and that the slope of the F0 contour is higher than neutral speech, especially at the end of the sentence. Paeschke *et al.* also analyzed the pitch shape in expressive speech [11]. They proposed different pitch features that might be useful for emotion recognition, such as the steepness of rising and falling of the pitch, and direction of the pitch contour [11]. Likewise, they also studied the differences in the global trend of the pitch, defined as the gradient of linear regression, in terms of emotions [12]. In all these experiments, they found statistically significant differences.

Bänziger and Scherer argued that the pitch mean and range account for most of the important emotional variation found in the pitch [13]. In our previous work, the mean, shape, and range of the pitch of expressive speech were systematically modified [14]. Then, subjective evaluations were performed to assess the emotional differences perceived in the synthesized sentences with the F0 modifications. The mean and the range were increased/decreased in different percentages and values. The pitch shape was modified by using stylization at varying semitone frequency resolution. The results indicated that modifications of the range (followed by the mean) had the biggest impact in the emotional perception of the sentences. The results also showed that the pitch shape needs to be drastically modified to change the perception of the original emotions. Using perceptual experiments, Ladd *et al.* also suggested that pitch range was more salient than pitch shape. Scherer *et al.* explained these results by making the distinction between linguistic and paralinguistic pitch features [15]. The authors suggested that gross statistics from the pitch are less connected to the verbal context, so they can be independently manipulated to express the emotional state of the speaker (paralinguistic). The authors also argued that the pitch shape (i.e., *rise* and *fall*) is tightly associated with the grammatical (linguistic) structure of the sentence.

Therefore, the pitch shape is jointly modified by linguistic and affective goals. As an aside, similar interplay with pitch has been observed in facial expressions [16].

Another interesting question is whether the emotional variations in the pitch contour change in terms of specific emotional categories or general activation levels. Bänziger and Scherer reported that the mean and range of the pitch contour change as a function of emotional arousal [13]. On the other hand, they did not find evidence for specific pitch shapes for different emotional categories. Thus, we argue that using pitch features is more suited for binary emotion classification than for implementing multiclass emotion recognition. These results support our ideas of contrasting pitch statistics derived from emotional speech with those of the neutral counterpart.

Although the aforementioned studies have reported statistically significant emotional differences, they do not provide automatic recognition experiments to validate the discriminative power of the proposed features. The framework presented in this paper allows us not only to identify the emotionally salient aspects the F0 contour, but also to quantify and compare their discriminative power for emotion recognition purposes. The main contributions of this paper are as follows:

- a discriminative analysis of emotional speech with respect to neutral speech;
- a novel methodology to analyze, quantify, and rank the most prominent and discriminative pitch features;
- a novel robust binary emotion recognition system based on contrasting expressive speech with reference neutral models.

III. Methodology

A. Overview

The fundamental frequency or F0 contour (pitch), which is a prosodic feature, provides the tonal and rhythmic properties of the speech. It predominantly describes the speech source rather than the vocal tract properties. Although it is also used to emphasize linguistic goals conveyed in speech, it is largely independent of the specific lexical content of what is spoken in most languages [17].

The fundamental frequency is also a supra-segmental speech feature, where information is conveyed over longer time scales than other segmental speech correlates such as spectral envelope features. Therefore, rather than using the pitch value itself, it is commonly accepted to estimate global statistics of the pitch contour over an entire utterance or sentence (sentence-level) such as the mean, maximum, and standard deviation. However, it is not clear that estimating global statistics from the pitch contour will provide local information of the emotional modulation [9]. Therefore, in addition to sentence-level analysis, we investigate alternative time units for the F0 contour analysis. Examples of time units that have been proposed to model or analyze the pitch contour include those at the foot-level [18], word-level [10], and even syllable-level [11]. In this paper, we propose to study the pitch features extracted over voiced regions (hereon referred as voiced-level). In this approach, the frames are labeled as voiced or unvoiced frames according to their F0 value (greater or equal to zero). Consecutive voiced frames are joined to form a voiced region over which the pitch statistics are estimated. The average duration of this time unit is 167 ms (estimated from the neutral reference corpus described in Section III-B). The lower and upper quartiles are 60 and 230 ms, respectively. The motivation behind using voiced region as a time unit is that the voicing process, which is influenced by the emotional modulation, directly determines voiced and unvoiced regions. Therefore, analysis along this level may shed further insights into emotional influence on the F0 contour not evident from the sentence level analyses. From a practical viewpoint, voiced regions are easier to segment compared to other short time units, which require forced alignment (word and syllable) or syllable stress detections (foot). In real-time applications, in which the audio is continuously recorded, this approach has the advantage that smaller buffers are required to process the audio. Also, it does not require pre-segmenting the input speech into utterances. Both sentence- and voiced-level pitch features are analyzed in this paper.

For the sake of generalization, the results presented in this paper are based on four different acted emotional databases (three for training and testing and one for validation) recorded from different research groups and spanning different emotional categories (Section III-B). Therefore, some degree of variability in the recording settings and the emotional elicitation is included in the analysis. Instead of studying the pitch contour in terms of emotional categories, the analysis is simplified to a binary problem in which emotional speech is contrasted with neutral speech (i.e., neutral versus emotional speech). This approach has the advantage of being independent of the emotional descriptors (emotional categories or attributes), and it is useful for many practical applications such as automatic expressive speech mining. In fact, it can be used as a first step in a more sophisticated multiclass emotion recognition system in which a second level classification would be used to achieve a finer emotional description of the speech.

Notice that the concept of neutral speech is not clear due to speaker variability. To circumvent this problem, we propose the use of a neutral (i.e., non-emotional) reference corpus recorded from many speakers (Section III-B). This neutral speech reference will be used to contrast the speech features extracted from the emotional databases (Section IV) to normalize the energy and the pitch contour for each speaker (Section III-C) and to build neutral model for emotional versus non-emotional classification (Section VII).

B. Databases

In this paper, five databases are considered: one non-emotional corpus used as a neutral speech reference, and four acted emotional databases with different properties. A summary of the databases is given in Table I.

The corpus considered in this paper as the neutral (i.e., non-emotional) reference database is the Wall Street Journal-based Continuous Speech Recognition Corpus Phase II (WSJ) [19]. This corpus, which we will refer to here on as WSJ1, comprises read and spontaneous speech from *Wall Street Journal* articles. For our purposes, only the spontaneous portion of this data was considered, which was recorded by 50 journalists with varying degrees of dictation experience. In total, more than eight thousand spontaneous utterances were recorded. Notice that in our

TABLE I
SUMMARY OF THE DATABASES (neu = neutral, ang = anger, hap = happiness, sad = sadness, bor = boredom, dis = disgust, fea = fear, anx = anxiety, pan = panic, anh = hot anger, anc = cold anger, des = despair, ela = elation, int = interest, sha = shame, pri = pride, con = contempt, sur = surprise)

Data	type	Use of the data	Spontaneous/acted	Language	# speakers	# utterances	Emotions
WSJ1	neutral	Reference	spontaneous	English	50	8104	neu
EMA	emotional	Training/testing	acted	English	3	688	neu,ang,hap,sad
EPSAT	emotional	Training/testing	acted	English	8	4738	neu,hap,sad,bor,dis,anx,pan,anh,anc,des,ela,int,sha,pri,con
GES	emotional	Training/testing	acted	German	10	535	neu,ang,hap,sad,bor,dis,fea
SES	emotional	Validation	acted	Spanish	1	266	neu,ang,hap,sad,sur

previous work [6], the read-speech TIMIT database was used as reference [20]. Since our ultimate goal is to build a robust neutral model for contrasting and recognizing emotion in real-life applications, this spontaneous corpus was preferred.

For the analysis and the training of the models (Sections IV–VI), three emotional corpora were considered. These emotional databases were chosen to span different emotional categories, speakers, genders, and even languages, with the purpose to include, to some extent, the variability found in the pitch. The first database was collected at the University of Southern California (USC) using an *electromagnetic articulography* (EMA) system [21]. In this database, which will be referred to here on as EMA, one male and two female subjects (two of them with formal theatrical vocal training) read ten sentences five times portraying the emotions sadness, anger, and happiness, in addition to neutral state. Although this database contains articulatory information, only the acoustic signals are analyzed in this study. Note that the EMA data have been perceptually evaluated and inter-evaluator agreement has been shown to be 81.9% [22].

The second emotional corpus corresponds to the Emotional Prosody Speech and Transcripts database (EPSAT) [23]. This database was collected at the University of Pennsylvania and is comprised of recordings from eight professional actors (five female and three male) who were asked to read short semantically neutral utterances corresponding to dates and numbers, expressing 14 emotional categories in addition to neutral state (Table I). The emotion states were elicited by providing examples of a situation for each emotional category. This database provides a wide spectrum of emotional manifestation.

The third emotional corpus is the Database of German Emotional Speech (GES) which was collected at the Technical University of Berlin [24]. This database was recorded from ten participants, five female, and five male, who were selected based on the naturalness and the emotional quality of the participant's performance in audition sessions. The emotional categories considered in the database are anger, happiness, sadness, boredom, disgust, and fear, in addition to neutral state. While the previous databases were recorded in English, this database was recorded in German. Although each language influences the shape of the pitch contour differently, we hypothesize that emotional pitch modulation can be still measured and quantified using English neutral pitch models. The assumption is that the fundamental frequency in English and German will share similar patterns. The results presented in Section VII-C give some evidence for this hypothesis.

In addition, a fourth emotional database is used in Section VII-C to evaluate the robustness of the pitch neutral models. Since the most discriminant F0 features are selected from the analysis presented in Sections IV and V, this database will be used to assess whether the emotional discrimination from this set of features extends to other corpora. This validation corpus corresponds to the Spanish Emotional Speech database (SES), which was collected from one professional actor with Castilian accent at the Polytechnic University of Madrid [25]. The emotions considered in this database are anger, happiness, sadness, and surprise, in addition to neutral state.

Although acted emotions differ from genuine emotions displayed during real-life scenarios [7], databases recorded from actors have been widely used in the analysis of emotions to cope with the inherent limitations of natural databases (e.g., copyright, lack of control, noisy signal). In fact, most of the current emotional corpora have been recorded from actors [26]. We have argued in our previous work about the role of acting as a viable research methodology for studying human emotions [27]. Although we acknowledge the simplifications of acted emotional speech, we believe that these corpora provide useful insights about the properties of genuine expressive speech.

C. Speaker Dependent Normalization

Normalization is a critical step in emotion recognition. The goal is to eliminate speaker and recording variability while keeping the emotional discrimination. For this analysis, a two-step approach is proposed: 1) energy normalization and 2) pitch normalization.

In the first step, the speech files are scaled such that the average RMS energy of the neutral reference database (E_{ref}) and the neutral subset in the emotional databases (E^s_{neu}) are the same for each speaker s. This normalization is separately applied for each subject in each database. The goal of this normalization is to compensate for different recording settings among the databases.

$$S^s_{\text{Energy}} = \sqrt{\frac{E_{\text{ref}}}{E^s_{\text{neu}}}}. \tag{1}$$

In the second step, the pitch contour is normalized for each subject (speaker-dependent normalization). The average pitch across speakers in the neutral reference database is estimated $F0_{\text{ref}}$. Then, the average pitch value for the neutral set of the emotional databases is estimated for each speaker $F0^s_{\text{neu}}$. Finally, a scaling factor (S^s_{F0}) is estimated by taking the ratio between $F0_{\text{ref}}$ and $F0^s_{\text{neu}}$, as shown in (2). Therefore, the neutral

TABLE II
SENTENCE- AND VOICED-LEVEL FEATURES EXTRACTED FROM THE F0

Description	Sentence level statistic		Voiced level statistic	
	F0	F0 derivative	F0	F0 derivative
F0 mean	Smean	Sdmean	Vmean	Vdmean
F0 standard deviation	Sstd	Sdstd	Vstd	Vdstd
F0 range	Srange	Sdrange	Vrange	Vdrange
F0 minimum	Smin	Sdmin	Vmin	Vdmin
F0 maximum	Smax	Sdmax	Vmax	Vdmax
F0 median	Smedian	Sdmedian	Vmedian	Vdmedian
F0 lower quartile	SQ25	SdQ25	VQ25	VdQ25
F0 upper quartile	SQ75	SdQ75	VQ75	VdQ75
F0 interquartile range	Siqr	Sdiqr	Viqr	Vdiqr
F0 kurtosis	Skurt	Sdkurt	**	**
F0 skewness	Sskew	Sdskew	**	**
F0 slope	**	**	Vslope	**
F0 curvature	**	**	Vcurv	**
F0 inflexion	**	**	Vinfle	**

samples of each speaker in the databases will have a similar F0 mean value

$$S^s_{F0} = \frac{F0_{\text{ref}}}{F0^s_{\text{neu}}}. \tag{2}$$

One assumption made in this two-step approach is that neutral speech will be available for each speaker. For real-life applications, this assumption is reasonable when either the speakers are known or a few seconds of their neutral speech can be prerecorded. Notice that these scaling factors will not affect emotional discrimination in the speech, since the differences in the energy and the pitch contour across emotional categories will be preserved.

D. Pitch Features

The pitch contour was extracted with the Praat speech processing software [28], using an autocorrelation method. The analysis window was set to 40 ms with an overlap of 30 ms, producing 100 frames per second. The pitch was smoothed to remove any spurious spikes by using the corresponding option provided by the Praat software.

Table II describes the statistics estimated from the pitch contour and the derivative of the pitch contour. These statistics are grouped into sentence-level and voiced-level features as defined in Section III-A. These are the statistics that are commonly used in related work to recognize emotions from the pitch. The nomenclature convention for the pitch features in this study was defined as intuitively as possible. Pitch features at sentence-level start with S. Pitch features at voiced-level start with V. The labels for the pitch derivative features start with either Sd (sentence-level), or Vd (voiced-level). Note that only voiced regions with more than four frames are considered to have reliable statistics (more than 40 ms). Likewise, kurtosis and skewness, in which the third and fourth moments about the mean need to be estimated, are not estimated at the voiced-level segments. As mentioned in Section III-A, the average duration of the voiced segments is 167 ms. (16.7 frames). Therefore, there are not enough samples to robustly estimate these statistics.

Describing the pitch shape for emotional modulation analysis is a challenging problem, and different approaches have been proposed. The *Tones and Break Indices System* (ToBI) is a well-known technique to transcribe prosody (or intonation) [29]. Although progress has been made toward automatic ToBI transcription [30], an accurate and more complete prosodic transcription requires hand labeling. Furthermore, linguistic models of intonation may not be the most appropriate labels to describe the emotions [13]. Taylor has proposed an alternative pitch contour parameterization called *Tilt Intonation Model* [31]. In this approach, the pitch contour needs to be pre-segmented into *intonation events*. However, there is no straightforward or readily available system to estimate these segments. Given these limitations, we follow a similar approach presented by Grabe *et al.* [32]. The voiced regions, which are automatically segmented from the pitch values, are parameterized using polynomials. This parameterization captures the local shape of the F0 contour with few parameters, which provides clear physical interpretation of the curves. Here, the slope (a_1), curvature (b_2), and inflexion (c_3) are estimated to capture the local shape of the pitch contour by fitting a first-, second-, and third-order polynomial to each voiced region segment

$$y = a_1 \cdot x + a_0 \tag{3}$$

$$y = b_2 \cdot x^2 + b_1 \cdot x + b_0 \tag{4}$$

$$y = c_3 \cdot x^3 + c_2 \cdot x^2 + c_1 \cdot x + c_0. \tag{5}$$

Table III shows additional sentence-level statistics derived from the voiced-level feature average. The nomenclature convention for these features is to start with SV. These statistics provide insights about the local dynamics of the pitch contour. For example, while the pitch range at the sentence-level (*Srange*) gives the extreme value distance of the pitch contour over the entire sentence, *SVmeanRange*, the mean of the range of the voiced regions, will indicate whether the voiced regions have flat or inflected shape. Likewise, some of these features will inform global patterns. For instance, the feature *SVmeanSlope* is highly correlated with the declination or global trend of the pitch contour, which previous studies have reported to convey emotional information [10], [12].

In sum, 60 pitch features are analyzed (39 sentence-level features and 21 voiced-level features). From here on, the statistics presented in Tables II and III are interchangeably referred to as "features," "F0 features," or "pitch features."

IV. EXPERIMENT 1: COMPARISONS USING SYMMETRIC KULLBACK–LEIBLER DISTANCE

This section presents our approach to identifying and quantifying the pitch features with higher levels of emotional modulation. Instead of comparing just the mean, the distributions of the pitch features extracted from the emotional databases are compared with the distributions of the pitch features extracted from the neutral reference corpus using KLD [33]. KLD provides a measure of the distance between two distributions. It is an appealing approach to robustly estimate the differences between the distributions of two random variables.

TABLE III
ADDITIONAL SENTENCE-LEVEL F0 FEATURES DERIVED FROM THE STATISTICS OF THE VOICED REGION PATTERNS

Global statistic derived from voiced segments	Value
Mean of the voiced segment ranges	SVmeanRange
Mean of the voiced segment maximums	SVmeanMax
Mean of the voiced segment minimums	SVmeanMin
Mean of the voiced segment lower quartiles	SVmeanQ25
Mean of the voiced segment upper quartiles	SVmeanQ75
Mean of the voiced segment interquartile ranges	SVmeanIqr
Mean of the voiced segment slopes	SVmeanSlope
Mean of the voiced segment curvatures	SVmeanCurv
Mean of the voiced segment inflexions	SVmeanInfle
Max. of the voiced segment slopes	SVmaxSlope
Max. of the voiced segment curvatures	SVmaxCurv
Max. of the voiced segment inflexion	SVmaxInfle
Max. of the voiced segment mean	SVmaxMean
Std. of the voiced segment means	SVstdMean
Std. of the voiced segment slopes	SVstdSlope
Std. of the voiced segment curvatures	SVstdCurv
Std of the voiced segment inflexions	SVstdInfle

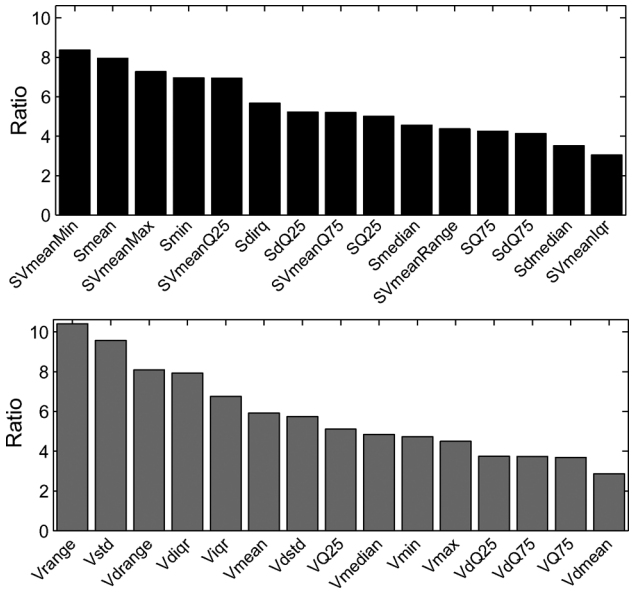


Fig. 1. Most emotional prominent features according to the average symmetric KLD ratio between features derived from emotional and neutral speech. The figures show the sentence-level (top) and voiced-level (bottom) features. The nomenclature of the F0 features is given in Tables II and III.

Since the KLD is not a symmetric metric, we propose the use of the symmetric Kullback–Leibler distance or $\mathcal{J}$-divergence, which is defined as

$$\mathcal{J}(q,p) = \frac{\mathcal{D}(q||p) + \mathcal{D}(p||q)}{2} \tag{6}$$

where $\mathcal{D}(p||q)$ is the conventional KLD

$$\mathcal{D}(q||p) = \sum_{\chi \in \mathcal{X}} q(\chi) \log \frac{q(\chi)}{p(\chi)}. \tag{7}$$

The first step is to estimate the distribution of the pitch features for each database, including the neutral reference corpus. For this purpose, we proposed the use of the K-means clustering algorithm to estimate the bins [34]. This nonparametric approach was preferred since the KLD is sensitive to the bins' estimation. To compare the symmetric KLD in terms of features and emotional categories k the number of bins, was set constant for each distribution ($k = 40$ empirically chosen). Notice that these feature-dependent nonuniform bins were estimated considering all the databases to include the entire range spanned by the features. After the bins were calculated, the distribution ($p_f^{(d,e)}$) of each pitch feature (f) was estimated for each database (d), and for each emotional category (e). Therefore, the true feature distribution for each subset is approximated by counting the number of samples assigned to each bin. The same procedure was used to estimate the distribution of the pitch features in the reference neutral corpus, q_f^{ref}.

The next step is to compute the symmetric KLD between the distribution of the emotional databases and the distribution estimated from the reference database $\mathcal{J}_f^{(d,e)}(p_f^{(d,e)}, q_f^{\text{ref}})$ (6). This procedure is repeated for each database and for each emotional category.

A good pitch feature for emotion discrimination ideally would have $\mathcal{J}_f^{(d,\text{neutral})}$ close to zero (neutral speech of the database d is similar to the reference corpus) and a high value for $\mathcal{J}_f^{(d,\bar{e})}$, where $\bar{e}$ is any emotional category except the neutral state. Notice that if $\mathcal{J}_f^{(d,\text{neutral})}$ and $\mathcal{J}_f^{(d,\bar{e})}$ have high values, this test would indicate that the speech from the emotional database is different from the reference database (*how neutral is the neutral speech?*). Likewise, if both values were similar, this feature would not be relevant for emotion discrimination. Therefore, instead of directly comparing the symmetric KLD, we propose to estimate the ratio between $\mathcal{J}_f^{(d,\bar{e})}$ and $\mathcal{J}_f^{(d,\text{neutral})}$ (8). That is, after matching the feature distributions with the reference feature distributions, the emotional speech is directly compared with the neutral set of the same emotional database by taking the ratio. High values of this ratio will indicate that the pitch features for emotional speech are different from their neutral counterparts, and therefore are relevant to discriminate emotional speech from neutral speech

$$r_f^{(d,\bar{e})} = \frac{\mathcal{J}_f^{(d,\bar{e})}}{\mathcal{J}_f^{(d,\text{neutral})}}. \tag{8}$$

Fig. 1 shows the average ratio between the emotional and neutral symmetric KLD obtained across databases and emotional categories. The pitch features with higher values are *SVmeanMin*, *SVmeanMax*, *Sdiqr*, and *Smean* for the sentence-level features and *Vrange*, *Vstd*, *Vdrange*, and *Vdiqr* for the voiced-level features. As further discussed in Section VI, these results indicate that gross statistics of the F0 contour are more emotionally salient than the features describing the pitch shape itself. In Section VII, the top features from this experiment will be used for binary emotion classification.

Figs. 2–4 show the results for the EMA, EPSAT, and GES databases, respectively. For the sake of space, these figures only display the results for the emotions anger, happiness, and sadness. They also include the average ratio across the emotional categories for each database (*Emo*). The figures

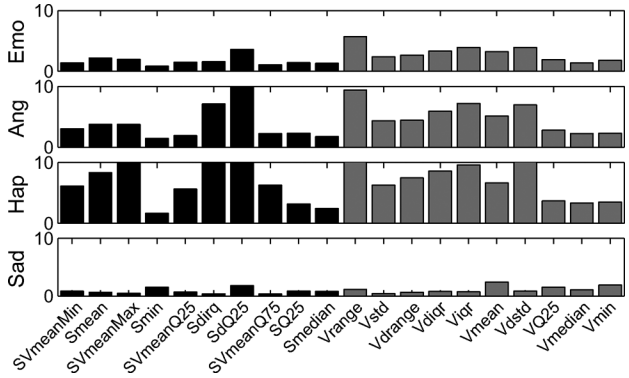


Fig. 2. Average symmetric KLD ratio between pitch features derived from emotional and neutral speech from the EMA corpus. The label *Emo* corresponds to the average results across all emotional categories. In order to keep the y axis fixed, some of the bars were clipped. The first ten bars correspond to sentence-level features and the last ten to voiced-level features. The nomenclature of the F0 features is given in Tables II and III.

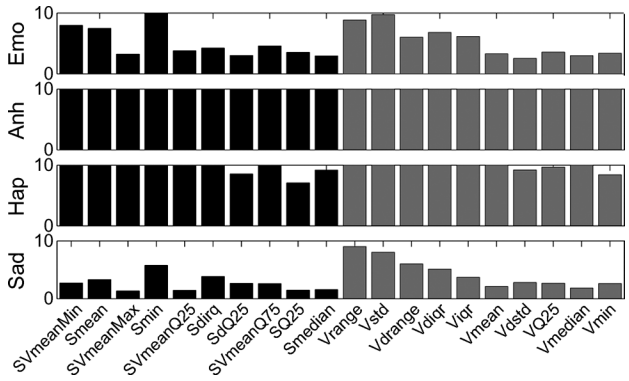


Fig. 3. Average symmetric KLD ratio between pitch features derived from emotional and neutral speech from the EPSAT corpus. The label *Emo* corresponds to the average results across all emotional categories. Only the emotional categories *hot anger*, *happiness*, and *sadness* are displayed. In order to keep the y axis fixed, some of the bars were clipped. The first ten bars correspond to sentence-level features and the last ten to voiced-level features. The nomenclature of the F0 features is given in Tables II and III.

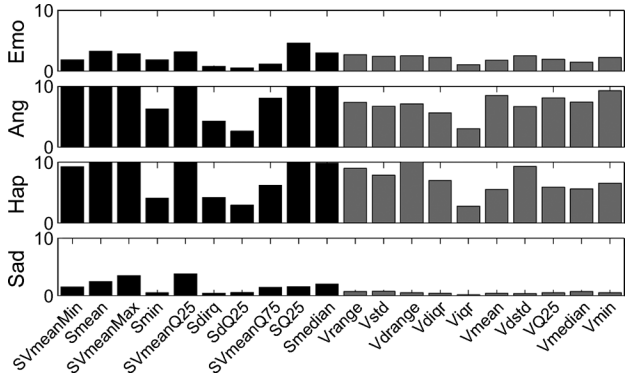


Fig. 4. Average symmetric KLD ratio between pitch features derived from emotional and neutral speech from the GES corpus. The label *Emo* corresponds to the average results across all emotional categories. Only the emotional categories *anger*, *happiness*, and *sadness* are displayed. In order to keep the y axis fixed, some of the bars were clipped. The first ten bars correspond to sentence-level features and the last ten to voiced-level features. The nomenclature of the F0 features is given in Tables II and III.

show that the rank of the most prominent pitch features varies according to the emotional databases. By analyzing different corpora, we hypothesize that the reported results will give more general insights about the emotional salient aspects of the fundamental frequency. These figures also reveal that some emotional categories with high activation levels (i.e., high arousal) such as anger and happiness are clearly distinguished from neutral speech using pitch-related features. However, subdued emotional categories such as sadness present similar pitch characteristics to neutral speech. This result agrees with the hypothesis that emotional pitch modulation is triggered by the activation level of the sentence [13], as mentioned in Section II. Further discussion about the pitch features is given in Section VI.

V. Experiment 2: Logistic Regression Analysis

The experiments presented in Section IV provide insight about the pitch features from expressive speech that differ from the neutral counterpart. However, they do not directly indicate the discriminative power of these features. This section addresses this question with the use of logistic regression analysis [35].

All the experiments reported in this section correspond to binary classification (neutral versus emotional speech). Unlike Section VII, the emotional databases are separately analyzed. The neutral reference corpus is not used in the section. The emotional categories are also separately compared with neutral speech (i.e., neutral-anger, neutral-happiness).

Logistic regression is a well-known technique to model binary or dichotomous variables. In this technique, the conditional expectation of the variable given the input variables is modeled with the specific form described in (9). After applying the *logit transformation* (10), the regression problem becomes linear in its parameters $(\beta_0, \ldots, \beta_n)$. A nice property of this technique is that the significance of the coefficients can be measured using the log-likelihood ratio test between two nested models (the input variables of one model are included in the other model). This procedure provides estimates about the discriminative power of each input feature

$$E(Y|f_1,\ldots,f_n) = \pi(x) = \frac{e^{\beta_0+\beta_1 x_1+,\ldots,\beta_n x_n}}{1+e^{\beta_0+\beta_1 x_1+,\ldots,\beta_n x_n}} \quad (9)$$

$$g(x) = \ln\left[\frac{\pi(x)}{1-\pi(x)}\right] = \beta_0+\beta_1 x_1+,\ldots,\beta_n x_n. \quad (10)$$

Experiment 2.1: The first experiment was to run logistic regression with only one pitch feature in the model at a time. The procedure is repeated for each emotional category. The goal of this experiment is to quantify the discriminative power of each pitch feature. This measure is estimated in terms of the improvement in the log-likelihood of the model when a new variable is added (the statistic $x = -2$ *log-likelihood ratio* is approximately chi-square distributed and can be used for hypothesis testing). Fig. 5 gives the average log-likelihood improvement across the emotional categories and databases for the top 15 sentence- and voiced-level features. The pitch features with higher score are *Smedian*, *Smean*, *SVmeanQ75*, and *SQ75* for the sentence-level features, and *VQ75*, *Vmean*, *Vmedian*, and *Vmax* for the voiced-level feature. These features will also be considered

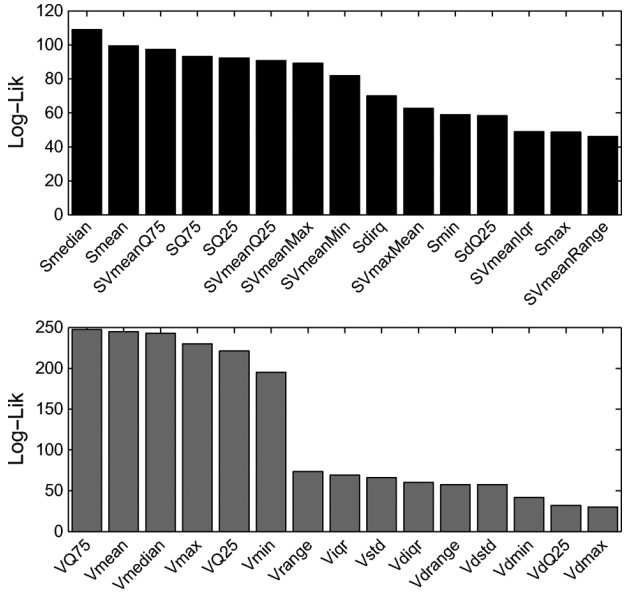


Fig. 5. Most emotionally discriminative pitch features according to the log-likelihood ratio scores in logistic regression analysis when only one feature is entered at a time in the models. The figure displays the average results across emotional databases and categories. The figures show the sentence-level (top) and voiced-level (bottom) features. The nomenclature of the F0 features is given in Tables II and III.

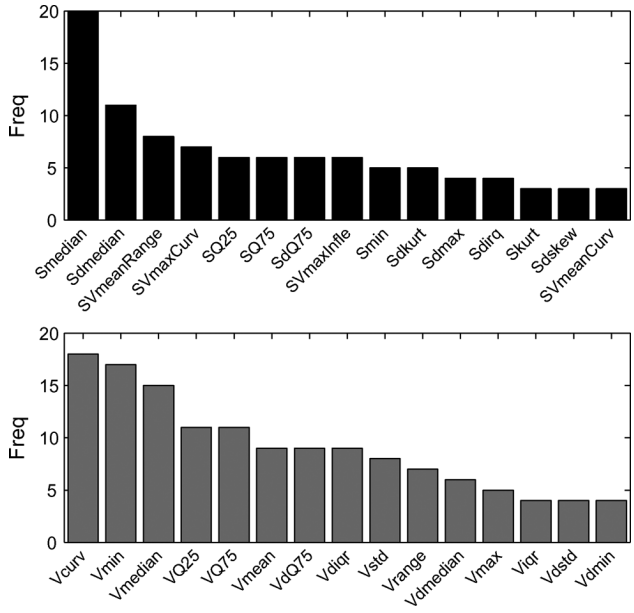


Fig. 6. Most frequently selected features in logistic regression models using forward feature selection. The figures show the sentence-level (top) and voiced-level (bottom) features. The nomenclature of the F0 features is given in Tables II and III.

for binary emotion recognition in Section VII. Although the order in the ranking in the F0 features is different in Figs. 1 and 5, eight sentence- and voiced-level features are included among the top ten features according to both criteria (experiments 1 and 2.1). This result shows the consistency of the two criteria used to identify the most emotionally salient aspects of the F0 contour (the F0 features with higher emotional/neutral symmetric KLD ratio are supposed to provide more discriminative information in the logistic regression models).

Experiment 2.2: Some of the pitch features provide overlapping information or are highly correlated. Since the pitch features were individually analyzed in experiment 2.1, these important issues were not addressed. Therefore, a second experiment was designed to answer this question, which is important for classification. Logistic regression analysis is used with *forward feature selection* (FFS) to discriminate between each emotional category and neutral state (i.e., neutral-anger). Here, the pitch features are sequentially included in the model until the log-likelihood improvement given the new variable is not significant (chi-square statistic test). In each case, the samples are split in training (70%) and testing (30%) sets.

Fig. 6 gives the pitch features that were most often selected in each of the 26 logistic regression tests (see Table IV). This figure provides insights about some pitch features, which may not be good enough if they are considered alone, but they give supplementary information to other pitch features. Notice that in each of these experiments, the pitch features were selected to maximize the performance of that specific task. The goal of analyzing the selected features across emotional categories and databases is to identify pitch features that can be robustly used to discriminate between emotional and neutral speech in a more general fashion (for generalization).

TABLE IV
DETAILS OF THE LOGISTIC REGRESSION ANALYSIS USING FFS WITH SENTENCE-LEVEL FEATURES (Acc = Accuracy, Rec = Recall, Pre = Precision, Bas = Baseline)

		Classification					Power of the model		
		Acc	Rec	Pre	F	Bas	-2 Log likelihood	Cox & Snell R^2	Nagelkerke R^2
EMA	Anger	91.9	92.0	93.9	92.9	58.1	66.6	0.675	0.901
	Happiness	96.0	95.3	98.4	96.8	64.0	16.6	0.729	0.976
	Sadness	68.2	63.5	78.6	70.2	59.1	314.7	0.124	0.166
	Emotional	87.1	91.0	93.2	92.1	82.2	370.3	0.341	0.491
GES	Anger	98.1	97.1	100	98.5	67.3	0.0	0.740	1.000
	Happiness	95.1	91.7	100	95.7	58.5	5.2	0.734	0.983
	Sadness	78.4	60.0	100	75.0	54.1	67.7	0.494	0.666
	Boredom	69.0	64.0	80.0	71.1	59.5	56.8	0.590	0.787
	Disgust	87.1	78.6	91.7	84.6	54.8	34.3	0.596	0.824
	Fear	92.9	88.0	100	93.6	59.5	25.5	0.671	0.903
	Emotional	83.1	93.0	88.7	90.8	89.4	249.2	0.191	0.324
EPSAT	Happiness	87.6	72.7	0.816	76.9	75.1	156.1	0.537	0.795
	Sadness	80.9	25.0	0.714	37.0	77.5	452.2	0.088	0.130
	Boredom	77.3	27.9	0.546	36.9	76.2	426.5	0.196	0.285
	Disgust	84.0	57.1	0.757	65.1	73.8	356.5	0.322	0.466
	Fear	74.7	15.9	0.438	23.3	75.8	480.1	0.133	0.192
	Panic	97.1	93.8	0.909	92.3	81.2	46.8	0.638	0.946
	Hot anger	97.1	97.1	0.895	93.1	79.8	37.7	0.635	0.954
	Cold anger	79.0	52.1	0.610	56.2	74.2	218.4	0.458	0.683
	Despair	79.3	50.0	0.732	59.4	69.7	276.4	0.368	0.553
	Elation	94.5	86.7	0.907	88.6	75.4	62.5	0.625	0.927
	Interest	85.6	68.4	0.796	73.6	70.8	254.2	0.430	0.634
	Shame	73.4	6.5	0.333	10.9	75.0	438.2	0.077	0.115
	Pride	83.3	54.8	0.677	60.5	76.7	318.5	0.296	0.446
	Contempt	76.4	33.3	0.704	45.2	70.8	465.0	0.148	0.214
	Emotional	80.5	97.1	0.824	89.2	82.6	1371.4	0.184	0.305

The pitch features that were most often selected in the logistic regression experiments reported in Fig. 6 are *Smedian*, *Sdmedian*, *SVmeanRange*, and *SVmaxCurv* for the sentence-level features, and *Vcurv*, *Vmin*, *Vmedian*, and *VQ25* for the

TABLE V
DETAILS OF THE LOGISTIC REGRESSION ANALYSIS USING FFS WITH VOICED-LEVEL FEATURES (Acc = Accuracy, Rec = Recall, Pre = Precision, Bas = Baseline)

		Classification					Power of the model		
		Acc	Rec	Pre	F	Bas	-2 Log likelihood	Cox & Snell R^2	Nagelkerke R^2
EMA	Anger	75.9	64.8	84.3	73.3	50.9	1730.7	0.320	0.426
	Happiness	83.2	74.5	91.2	82.0	51.4	1429.9	0.424	0.565
	Sadness	59.5	58.4	60.9	59.6	51.2	2371.0	0.043	0.057
	Emotional	76.5	96.6	77.8	86.2	75.9	3286.9	0.167	0.246
GES	Anger	91.0	91.5	93.6	92.6	61.3	581.5	0.541	0.751
	Happiness	88.0	81.8	94.4	87.7	52.1	565.6	0.477	0.636
	Sadness	59.9	61.5	61.5	61.5	52.1	1014.8	0.105	0.140
	Boredom	55.7	42.6	58.6	49.3	50.6	1026.8	0.041	0.054
	Disgust	73.1	43.0	76.7	55.1	61.6	761.1	0.174	0.236
	Fear	88.7	83.5	90.6	86.9	55.3	570.6	0.458	0.612
	Emotional	85.6	99.4	86.1	92.3	86.1	2002.2	0.100	0.183
EPSAT	Happiness	82.3	43.8	85.4	57.9	72.2	1120.6	0.278	0.413
	Sadness	74.7	10.6	66.7	18.3	73.3	1487.1	0.044	0.065
	Boredom	70.6	0.0	0.0	0.0	71.3	1594.8	0.015	0.023
	Disgust	72.4	22.0	70.0	33.5	68.5	1558.7	0.129	0.186
	Fear	72.0	2.5	44.4	4.7	72.2	1677.9	0.024	0.034
	Panic	88.6	62.3	86.2	72.3	76.1	581.3	0.457	0.707
	Hot anger	91.3	74.5	88.7	81.0	75.2	720.3	0.451	0.672
	Cold anger	77.4	23.7	75.0	36.0	73.2	1340.4	0.164	0.243
	Despair	77.1	26.3	71.9	38.5	72.7	1368.5	0.146	0.216
	Elation	89.8	71.2	90.6	79.7	71.8	719.4	0.446	0.668
	Interest	77.4	33.5	80.0	47.2	69.9	1276.6	0.205	0.303
	Shame	76.7	0.0	0.0	0.0	76.7	1530.3	0.010	0.015
	Pride	76.9	14.9	65.8	24.2	74.8	1332.7	0.133	0.200
	Contempt	75.3	17.1	65.0	27.1	73.2	1576.1	0.077	0.112
	Emotional	83.8	99.9	83.9	91.2	83.9	4980.5	0.101	0.167

voiced-level features. This experiment reveals interesting results. For example, *Smedian* and *Smean* were seldom selected together since they are highly correlated ($\rho \approx 0.96$). In fact, while *Smean* was the second best feature according to the experiment 2.1 (Fig. 5), it is not even in the top 15 features according to this criterion (Fig. 6). In contrast, other features that were not relevant when they were individually included in the model appear to provide important supplementary information (e.g., *SVmeanCurv* and *Vcurv*). Further discussion about the features is given in Section VI.

Tables IV and V provide details about the logistic regression experiments performed with FFS for the sentence- and voiced-level features, respectively. In the tables, we highlight the cases when the fit of the logistic regression models was considered adequate, according to the Nagelkerke r-square statistic ($r^2 > 0.4$, empirically chosen) [36]. These tables show that some emotional categories cannot be discriminated from neutral speech based on these pitch features (e.g., sadness, boredom, shame). The tables also reveal that voiced-level features provide emotional information, but the performance is in general worse than the sentence-level features. This result indicates that the voiced segment region may not be long enough to capture the emotional information. An alternative hypothesis is that not all the voiced region segments present measurable emotional modulation, since the emotion is not uniformly distributed in time [37]. In fact, previous studies suggest that the patterns displayed by the pitch at the end of the sentences are important for emotional categories such as happiness [10], [15]. Therefore, the confusion in the classification task may increase by considering each voiced region as a sample.

VI. ANALYSIS OF PITCH FEATURES

On the one hand, the results presented in the previous sections reveal that pitch statistics such as the mean/median, maximum/upper quartile, minimum/lower quartile, and range/interquartile range, are the most emotionally salient pitch features. On the other hand, features that describe the pitch contour shape such as the slope, curvature and inflexion, are not found to convey the same measurable level of emotional modulation. These results indicate that the continuous variations of pitch level are the most salient aspects that are modulated in expressive speech. These results agree with previous findings reported in [13] and [38], which indicate that pitch global statistics such as the mean and range are more emotionally prominent than the pitch shape itself, which is more related with the verbal context of the sentence [15].

The results of the experiment 1 indicate that the standard deviation and its derivative convey measurable emotional information at the voiced-level analysis (*Vstd*, Fig. 1). This result agrees with the finding reported by Lieberman and Michaels, which suggested that fluctuations in short-time segments are indeed important emotional cues [9]. Notice that in the experiments 2.1 and 2.2 reported in Section V, *Vstd* is among the top-ten best features (Figs. 5 and 6).

The results in Fig. 6 suggest that the curvature of the pitch contour is affected during expressive speech. Although *SVmaxCurv* and *Vcurv* were never selected as the first feature in the FFS algorithm, they are among the most selected features for the sentence- and voiced-level logistic regression experiments. These results indicate that these features provide supplementary emotional information that can be used for classification purposes. For other applications such as expressive speech synthesis, changing the curvature may not significantly change the emotional perception of the speech. This result agrees with the finding reported by Bulut and Narayanan [14] (Section II).

The analysis also reveals that sentence-level features derived from voiced segment statistics (Table III) are important. From the top-five sentence-level features in Figs. 1, 5, and 6, six out of twelve features correspond to global statistics derived from voiced segments. This result suggests that variations between voiced regions convey measurable emotional modulation.

Features derived from the pitch derivative are not as salient as the features derived from the pitch itself. Also, *SVmeanSlope*, which is related to the pitch global trend, is not found to be an emotionally salient feature, as suggested by Wang *et al.* and Paeschke [10], [12].

To build the neutral models for binary emotion recognition (Section VII), a subset of the pitch features was selected. Instead of finding the best features for that particular task, we decided to pre-select the top-six sentence- and voiced-level features based on results from experiments 1, 2.1 and 2.2 presented in Sections IV and V (Figs. 1, 5, and 6). Some of the features were removed from the group since they presented high levels of correlation. The pitch features *Sdiqr*, *Smedian*, *SQ75*, *SQ25*, *Sdmedian*, *SVmeanRange*, and *SVmaxCurv* were selected as sentence-level features, and *Vstd*, *Vdrange*, *Vdiqr*, *VQ75*, *Vmedian*, *Vmax*, and *Vcurv* were selected as voiced-level features. Table VI gives the correlation matrix between these features.

TABLE VI
CORRELATION OF THE SELECTED PITCH FEATURES

	Sentence-level features						
	Sdiqr	Smedian	SQ75	SQ25	Sdmedian	SVmeanRange	SVmaxCurv
Sdiqr	1.000	**0.709**	**0.751**	0.641	-0.211	0.668	-0.107
Smedian	**0.709**	1.000	**0.897**	**0.956**	-0.268	0.520	-0.227
SQ75	**0.751**	**0.897**	1.000	**0.834**	-0.252	0.575	-0.176
SQ25	0.641	**0.956**	**0.834**	1.000	-0.248	0.455	-0.224
Sdmedian	-0.211	-0.268	-0.252	-0.248	1.000	-0.166	0.098
SVmeanRange	0.668	0.520	0.575	0.455	-0.166	1.000	0.178
SVmaxCurv	-0.107	-0.227	-0.176	-0.224	0.098	0.178	1.000
	Voiced-level features						
	Vstd	Vdrange	Vdiqr	VQ75	Vmedian	Vmax	Vcurv
Vstd	1.000	**0.910**	**0.789**	0.477	0.231	0.656	0.152
Vdrange	**0.910**	1.000	**0.800**	0.491	0.291	0.685	0.053
Vdiqr	**0.789**	**0.800**	1.000	0.617	0.428	0.664	-0.082
VQ75	0.477	0.491	0.617	1.000	**0.952**	**0.937**	-0.394
Vmedian	0.231	0.291	0.428	**0.952**	1.000	0.855	-0.525
Vmax	0.656	0.685	0.664	**0.937**	0.855	1.000	-0.276
Vcurv	0.152	0.053	-0.082	-0.394	-0.525	-0.276	1.000

Only a few pitch features present high levels of correlation. These variables were not removed since our preliminary results indicated that they improve the recognition accuracy.

VII. EMOTIONAL DISCRIMINATION RESULTS USING NEUTRAL MODELS

In this section, neutral speech prosodic models are trained for emotional speech discrimination. Aspects of this approach were originally proposed to analyze the emotional modulation observed in expressive speech [39]. In our recent study, we proposed this framework to recognize expressive speech using the acoustic likelihood scores obtained from *hidden Markov models* (HMMs) [6]. The models were trained with neutral (non-emotional) speech using spectral features. In this section, the ideas are extended to build neutral models for the selected sentence- and voiced-level pitch features (Table VI).

A. Motivation and Proposed Approach

Automatic emotion recognition in real-life applications is a nontrivial problem due to the inherent inter-speaker variability of expressive speech. Furthermore, the emotional descriptors are not clearly established. The boundaries between emotional categories are blurred [7] and do not account for different degrees of emotional intensity [40]. Most of the current efforts to address this problem have been limited to dealing with emotional databases spanning a subset of emotional categories. The feature selection and the models are trained for specific databases with the risk of sparseness in the feature space and over-fitting. It is also fairly difficult, if not infeasible, to collect enough emotional speech data so that one can train robust and universal acoustic models of individual emotions. Therefore, it is not surprising that the models built with these individual databases (usually offline) do not easily generalize to different databases or online recognition tasks in which blending of emotions is observed [26].

Instead of building emotional models, we propose the use of robust acoustic neutral reference models to discriminate emotional speech, under the assumption that expressive speech differs from neutral speech in the measurable feature space. One advantage of this approach is that many more emotionally neutral databases are available to build robust models. Since we are addressing the problem of neutral versus emotional speech, this approach does not depend on the emotional labels used to tag the corpus. Furthermore, the framework inherently captures speaker variability; for our experiments, the reference models are built with the WSJ1 database (Section III-B), which was collected from 50 speakers.

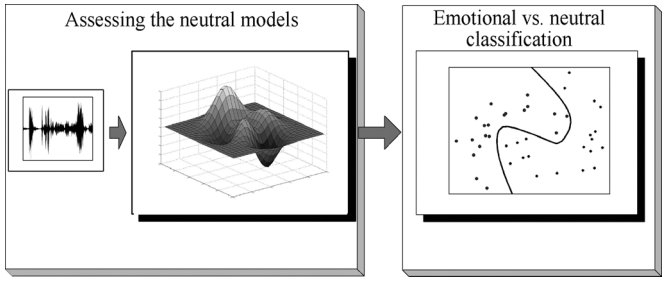


Fig. 7. General framework of the proposed two-step approach to discriminate neutral versus emotional speech. In the first step, the input speech is contrasted with robust neutral references models. In the second step, the *fitness measures* are used for binary emotional classification (details are given in Section VII-A). In this paper, the neutral models are implemented with univariate GMMs (for each F0 feature), and the classification is implemented with LDC.

Fig. 7 describes the general framework of the proposed two-step approach. In the first step, neutral models are built to measure the degree of similarity between the input speech and the reference neutral speech. The output of this block is a *fitness measure* of the input speech. In the second step, these measures are used as features to infer whether the input speech is emotional or neutral. If the features from the expressive speech differ in any aspect from their neutral counterparts, the fitness measure will decrease. Therefore, we hypothesize that setting thresholds over these *fitness measures* is easier and more robust than setting thresholds over the features themselves.

While the first step is independent of the emotional database, the speakers, and the emotional categories, the second step depends on these factors since the classifier needs to be trained with emotional and neutral speech. To overcome this limitation, the three emotional databases (EMA, EPSAT, and GES) were combined to train a semi-corpus-independent classifier. Notice that this binary recognition task is more challenging than the logistic regression analysis presented in Section V, since the emotional corpora are jointly used, and all the emotional categories (without neutral state) are grouped into a single category (emotional).

The proposed two-step framework described in Fig. 7 is general and can be implemented using different algorithms. For example, in our previous work, we built neutral models (first step) for spectral features using HMMs [6]. These models were dependent on the underlying phonetic units of the spoken message. Likewise, any linear or nonlinear machine learning technique can be used to classify the *fitness measures* (second step). The proposed approach can be extended to other features such as voice quality or even facial features (i.e., comparing neutral faces with expressive faces).

As mentioned in Section III-A, the F0 contour is assumed to be largely independent of the specific lexical content, in contrast

to spectral speech features. Therefore, a single lexical-independent model is adequate to model the selected pitch features. For this task, we propose the use of univariate GMM for each pitch feature f

$$\mathcal{F}_f(X_f = \chi_f|\Theta) = \sum_{j=1}^{K} \alpha_j \frac{1}{\sigma_j\sqrt{2\pi}} \exp\left(\frac{(\chi_f - \mu_j)^2}{-2\sigma_j^2}\right) \quad (11)$$

with

$$\Theta = \{\alpha_j, \mu_j, \sigma_j\}_{j=1}^{K} \quad \alpha_j > 0 \; j = 1, \ldots, K, \; \sum_{j=1}^{K} \alpha_j = 1.$$

The maximum likelihood estimates of the parameters in the GMM Θ are computed using the *expectation-maximization* (EM) algorithm. These parameters are estimated with the pitch features derived from the WSJ1 corpus (reference neutral database). For initialization, k samples are selected at random with uniform mixing proportions (α). The maximum number of iteration was set to 200.

For a given input speech, the likelihoods of the models, $\mathcal{F}_f(X_f = \chi_f|\Theta)$, are used as *fitness measures*. In the second step, a *Linear Discriminant Classifier* (LDC) was implemented to discriminate between neutral and expressive speech. While more sophisticated non-linear classifiers may give better accuracy, this linear classifier was preferred for the sake of generalization.

B. Results

The recognition results presented in this section are the average values over 400 realizations. Since the emotional categories are grouped together, the number of emotional samples is higher than the neutral samples. Therefore, in each of the 400 realizations, the emotional samples were randomly drawn to match the number of neutral samples. Thus, for the experiments presented here and in Section VII-C, the priors were equally set for the neutral and emotional classes $(\text{baseline} = 50\%)$. Then, the selected samples were split in training and testing sets (70% and 30%, respectively). Notice that the three emotional corpora are considered together.

Given that some emotional categories are confused with neutral speech in the pitch feature space (Section V), a subset of emotional categories for each database was selected. The criterion was based on the Nagalkerke r-square score of the logistic regression presented in Table IV $(R^2 > 0.4)$. This section presents the results in terms of all emotional categories (*all emotions*) and this subset of emotional categories (*selected emotions*).

An important parameter of the GMM is the number of mixtures, K. Fig. 8 shows the performance of the GMM-based pitch neutral models for different numbers of mixtures. The figure shows that the proposed approach is not sensitive to this parameter. For the rest of the experiments, K was set to 2.

Table VII presents the performance of the proposed approach for the sentence- and voiced-level features. When all the emotional categories are used, the performance accuracy reaches 77.31% for the sentence-level features and 72% for the voiced-level features. These values increase approximately 5% when only the selected emotional categories are considered. Notice that only pitch features are used, so these values are notably high compared to the baseline (50%).

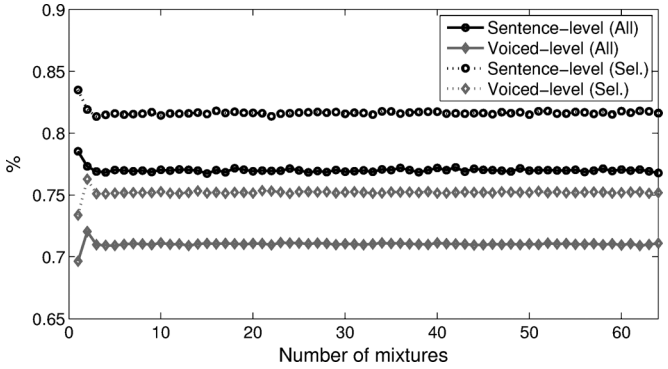


Fig. 8. Accuracy of the proposed neutral model approach as a function of the number of mixture components in the GMM. The results are not sensitive to this variable. For the rest of the experiments $k = 2$ was selected.

TABLE VII
ACCURACY OF THE PROPOSED NEUTRAL MODEL APPROACH AS A FUNCTION OF THE FEATURE TYPE AND EMOTION SET. THE ACCURACIES OF THE CONVENTIONAL LDC CLASSIFIER (WITHOUT NEUTRAL MODELS) FOR THE SAME TASK ARE ALSO PRESENTED

	Neutral Model		Conventional scheme	
	sentence	voiced	sentence	voiced
All emotions	77.31%	72.00%	74.67%	70.96%
Selected emotions	81.88%	76.19%	80.36%	75.42%

For comparison, Table VII also presents the results for the same task, using the pitch statistics as features without the neutral models (without the first step in the proposed approach as described in Fig. 9). This classifier, which is similar to the conventional frameworks used to discriminate emotions, was also implemented with LDC. The table shows that the proposed approach achieves better performance than the conventional approach in each of the four conditions (sentence/voiced level features; all/selected emotional categories). A paired samples t-test was computed over the 400 realizations to measure whether the differences between these two approaches are statistically significant. The results indicate that the classifier trained with the likelihood scores (proposed approach) is significantly better than the one trained with the F0 features (using the conventional approach) in each of the four conditions $(p \ll 0.001)$. In Section VII-C, the neutral model approach is compared with the conventional LDC classifier in terms of robustness.

In Table VIII, the results of the proposed approach are disaggregated in terms of databases (notice that three different emotional databases are used for training and testing). An interesting result is that the precision rate is in general high, which means that there are not many neutral samples labeled as emotional (false positive). For the sentence-level features, the accuracy for the EPSAT database is slightly lower than for the other databases (6%–11%). This result might be explained by the short sentences used to record this corpus (Section III-B).

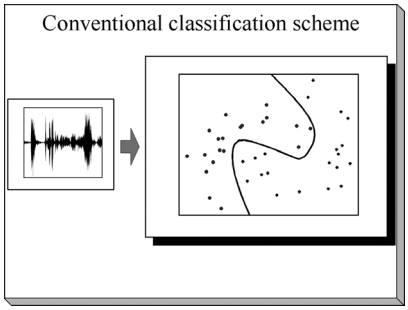


Fig. 9. Conventional classification scheme for automatic emotion recognition. Speech features are directly used as input of the classifier, instead of the fitness measures estimated from the neutral reference models (Fig. 7). The classifier is implemented with LDC.

TABLE VIII
PERFORMANCES OF THE PROPOSED NEUTRAL MODEL APPROACH FOR EACH EMOTIONAL DATABASE (Acc = Accuracy, Rec = Recall, Pre = Precision)

	Sentence level (all emotions)				Voiced level (all emotions)			
	Acc	Rec	Pre	F	Acc	Rec	Pre	F
EMA	0.865	0.726	0.921	0.812	0.742	0.622	0.707	0.662
GES	0.809	0.779	0.867	0.821	0.711	0.688	0.799	0.739
EPSAT	0.740	0.733	0.757	0.745	0.710	0.623	0.777	0.691
	Sentence level (selected emotions)				Voiced level (selected emotions)			
	Acc	Rec	Pre	F	Acc	Rec	Pre	F
EMA	0.905	0.808	0.942	0.870	0.797	0.716	0.740	0.727
GES	0.799	0.760	0.916	0.831	0.703	0.670	0.854	0.751
EPSAT	0.798	0.805	0.792	0.798	0.773	0.727	0.790	0.757

With few frames available to estimate the pitch statistics, the F0 features will have higher variability and will be less accurate.

Table IX provides further details about the recognition performance of the proposed approach. In this table, the results are disaggregated in terms of the emotional categories for each database. The results are presented in terms of recall rate (accuracy and precision values are given in Table VIII). In most of the cases, the recall rate is equal to or better than the recall rates reported in the logistic regression experiments (Tables IV and V). Notice that this task is significantly harder than the task presented in Section V, since the emotional categories and the emotional database were jointly analyzed.

C. Robustness of the Neutral Model Approach

As mentioned before, using neutral models for emotional recognition is hypothesized to be more robust and, therefore, to generalize better than using a direct emotion classification approach. To validate this claim, this section compares the performance of the proposed approach (Fig. 7) with the conventional classifier (without neutral models, Fig. 9) when there is a mismatch between the training and testing conditions. For this purpose, the emotional databases were separated by languages into two groups: English (EPSAT, EMA), and German (GES). One of these groups was used for training, and the other one for testing. The results for the two conditions are given in Table X for the sentence-level features and Table XI for the voiced-level features. Since the samples were randomly drawn to have an equal number of emotional and neutral samples (both in the training and testing sets), the baseline is 50%. The recognition results reported here are also average values over 400 realizations.

TABLE IX
RECALL RATE OF THE PROPOSED NEUTRAL MODEL APPROACH FOR EACH EMOTIONAL CATEGORY (Rec = Recall)

		All emotions		Selected emotions	
		sentence Rec	voiced Rec	sentence Rec	voiced Rec
EMA	Anger	0.739	0.668	0.710	0.654
	Happiness	0.917	0.787	0.907	0.777
	Sadness	0.520	0.421	***	***
GES	Anger	1.000	0.953	1.000	0.947
	Happiness	0.961	0.887	0.968	0.875
	Sadness	0.463	0.214	0.427	0.189
	Boredom	0.333	0.340	0.226	0.306
	Disgust	0.891	0.698	0.866	0.665
	Fear	0.920	0.859	0.932	0.848
EPSAT	Happiness	0.881	0.792	0.875	0.785
	Sadness	0.555	0.391	***	***
	Boredom	0.424	0.338	***	***
	Disgust	0.610	0.545	0.606	0.534
	Fear	0.666	0.436	***	***
	Panic	0.982	0.971	0.980	0.968
	Hot anger	0.987	0.945	0.986	0.945
	Cold anger	0.778	0.648	0.760	0.632
	Despair	0.747	0.656	0.724	0.646
	Elation	0.963	0.928	0.962	0.924
	Interest	0.738	0.659	0.727	0.648
	Shame	0.545	0.425	***	***
	Pride	0.724	0.563	0.697	0.541
	Contempt	0.730	0.526	***	***

For sentence-level F0 features, Table X shows that the neutral model approach generalizes better than the conventional scheme. In fact, the absolute accuracy improvement over the conventional scheme is over 4%. Even though there is a mismatch between the training and testing conditions, the performance of the proposed approach does not decrease compared to the case when the same corpora are used for training and testing (no mismatch). For instance, Table VIII shows that the accuracy of the GES database was 80.9% when there was not a training/testing mismatch. Interestingly, Table X shows that the performance for this database is still over 80% when only the English databases are used for training. When the classifier is trained with the German database, and tested with the English databases, the performance is 75.1%. As mentioned in Section VII-B, the EPSAT database presents the lowest performance of the emotional databases considered in this paper (74%, Table VIII). Since this corpus accounts for more than 85% of the English samples (Table I), the lower accuracy observed for the English databases is expected.

For the voiced-level F0 features, Table XI shows that the performance of the proposed approach is similar to the performance of the system without any mismatch (see Table VIII). The conventional scheme presents similar performance.

Notice that the F0 features were selected from the analysis presented in Sections IV and V. The EMA, EPSAT, and GES databases were considered for the analysis. To assess whether

TABLE X
VALIDATING THE ROBUSTNESS OF THE NEUTRAL MODEL APPROACH AGAINST MISMATCH BETWEEN TRAINING AND TESTING CONDITIONS. SENTENCE-LEVEL FEATURES (Acc = Accuracy, Rec = Recall, Pre = Precision)

Databases		Neutral model				Conventional scheme				ΔAcc
Training	Testing	Acc	Rec	Pre	F	Acc	Rec	Pre	F	
English (EPSAT,EMA)	German (GES)	0.802	0.778	0.818	0.798	0.761	0.620	0.864	0.722	4.1%
German (GES)	English (EPSAT,EMA)	0.751	0.732	0.762	0.746	0.705	0.509	0.837	0.633	4.6%
English (EPSAT,EMA)	Spanish (SES)	0.782	0.739	0.809	0.772	0.604	0.412	0.668	0.510	17.9%
German (GES)	Spanish (SES)	0.792	0.708	0.851	0.773	0.686	0.445	0.857	0.586	10.6%
English, German (EPSAT,EMA,GES)	Spanish (SES)	0.794	0.729	0.838	0.780	0.649	0.420	0.775	0.545	14.5%

TABLE XI
VALIDATING THE ROBUSTNESS OF THE NEUTRAL MODEL APPROACH AGAINST MISMATCH BETWEEN TRAINING AND TESTING CONDITIONS. VOICED-LEVEL FEATURES (Acc = Accuracy, Rec = Recall, Pre = Precision)

Databases		Neutral model				Conventional scheme				ΔAcc
Training	Testing	Acc	Rec	Pre	F	Acc	Rec	Pre	F	
English (EPSAT,EMA)	German (GES)	0.710	0.695	0.716	0.705	0.733	0.590	0.827	0.689	-2.4%
German (GES)	English (EPSAT,EMA)	0.716	0.594	0.787	0.677	0.699	0.520	0.809	0.633	1.8%
English (EPSAT,EMA)	Spanish (SES)	0.681	0.564	0.736	0.639	0.641	0.333	0.868	0.481	4.0%
German (GES)	Spanish (SES)	0.692	0.518	0.794	0.627	0.634	0.314	0.872	0.462	5.8%
English, German (EPSAT,EMA,GES)	Spanish (SES)	0.684	0.555	0.749	0.638	0.641	0.328	0.876	0.477	4.4%

the emotional discrimination observed from these F0 features transpires to other corpora, a fourth emotional database was considered for the final experiments. For this purpose, the SES database is used, which was recorded in Spanish (Section III-B). Notice than the SES corpus contains the emotional category *surprise*, which is not included in the training set.

For this experiment, the classifier of the neutral model approach was separately trained with the English (EPSAT, EMA), German (GES), and English and German databases. Tables X and XI present the results for the sentence- and voiced-level F0 features, respectively. The results indicate that the accuracy of the proposed approach is over 78% for the sentence-level features and 68% for the voiced level features. The performance is similar to the ones achieved with the other emotional databases considered in this paper. Interestingly, the performance of the proposed approach is about 10%–18% (absolute) better than the one obtained with the conventional scheme. These results suggest that conventional approaches to automatically recognizing emotions are sensitive to the feature selection process (the most discriminant features from one database may not have the same discriminative power in another corpus). However, the performance of the proposed approach can be robust against this type of variability.

In Section III-B, we hypothesized that neutral speech prosodic models trained with English speech can be used to detect emotional speech in another language. The results presented in Tables X and XI support this hypothesis. As mentioned in Sections III-A, the fundamental frequency in language such as German, English, and Spanish is largely independent of the specific lexical content of the utterance. As a result, the proposed neutral speech prosodic models present similar performance regardless of the languages of the databases used to train and test the classifier.

VIII. CONCLUSION

This paper presented an analysis of different expressive pitch contour statistics with the goal of finding the emotionally salient aspects of the F0 contour (pitch). For this purpose, two experiments were proposed. In the first experiment, the distribution of different pitch features was compared with the distribution of the features derived from neutral speech using the symmetric KLD. In the second experiment, the emotional discriminative power of the pitch features was quantified within a logistic regression framework. Both experiments indicate that dynamic statistics such as mean, maximum, minimum, and range of the pitch are the most salient aspects of expressive pitch contour. The statistics were computed at sentence and voiced region levels. The results indicate that the system based on sentence-level features outperforms the one with voiced-level statistics both in accuracy and robustness, which facilitates a turn-by-turn processing in emotion detection.

The paper also proposed the use of neutral models to contrast expressive speech. Based on the analysis of the pitch features, a subset with the most emotionally salient features was selected. A GMM for each of these features was trained using a reference neutral speech corpus (WSJ1). After contrasting the input speech with neutral models, the likelihood scores were used for classification. The approach was trained and tested with three different emotional databases spanning different emotional categories, recording settings, speakers, and even languages (English and German). The recognition accuracy of the proposed approach was over 77% (baseline 50%) using only pitch-related features. To validate the robustness of the approach, the system was trained and tested with different databases recorded in three different languages (English, German, and Spanish). Although there was a mismatch between the training and testing condition,

the performance of the proposed framework did not degrade. In contrast, the performance of the conventional classifier without the neutral models decreased up to 17.9% (absolute, Table X), for the same task using the same F0 features. These results show that this system is robust against different speakers, languages, and emotional descriptors and can generalize better than standard emotional classifiers.

Results from our previous work indicated that emotional modulation is not uniformly distributed, in time and in space, across different communicative channels [37]. If this trend is also observed in the fundamental frequency, certain regions in the pitch contour might present stronger emotional modulation, as discussed in Section V. We are planning to study this problem by comparing neutral and emotional utterances spoken with the same lexical content. With this approach, we would be able to locally compare the F0 contours between emotional and neutral speech under similar lexical constraints.

As mentioned in Section III-A, the proposed approach to detect emotional speech can be used as a first step in a multiclass emotion recognition system. In many domains, neutral speech is more common than expressive speech (e.g., call centers). Therefore, it is very useful to have a robust emotional speech detector at the front end. Depending on the application, the emotional speech can be postprocessed using emotion specific models. For example, in call center applications, the emotional speech could be further classified as positive or negative based on activation specific models.

One drawback of the emotional databases used in this paper is that they were collected from actors. In our future work, we will include in the analysis natural emotional databases recorded from real-life scenarios (e.g., the Vera am Mittag German audio-visual emotional speech database [41]).

Another limitation of this framework is that speaker dependent normalization is used to reduce speaker variability. In general, neutral speech for each speaker may not be available. In that scenario, at least gender normalization should be applied [42]. Our ultimate goal is to design a framework to automatically detect emotional speech regions from large amounts of data in an unsupervised manner (e.g., call center data). Therefore, we are currently working on extending the proposed approach by using speaker-independent normalization.

In terms of classification, we are planning to expand the proposed approach to include features related to energy and duration. Likewise, this neutral prosodic model can be combined with the neutral spectral models presented in our previous work [6]. By considering different emotionally salient aspects of speech, we expect to improve the accuracy and robustness of the proposed neutral model approach further.

ACKNOWLEDGMENT

The authors would like to thank the anonymous reviewers for their thoughtful and insightful comments. The authors would also like to thank colleagues in the SAIL emotion research group for their valuable comments. Special thanks go to M. Bulut, J. Tepperman, M. Francis, and A. Kazemzadeh. The SES database is a property of Universidad Politécnica de Madrid, Departamento de Ingeniería Electrónica, Grupo de Tecnología del Habla, Madrid, Spain, and their help is gratefully acknowledged. Any opinions, findings, and conclusions or recommendations expressed in this paper are those of the authors and do not necessarily reflect the views of the funding agencies.

REFERENCES

[1] R. W. Picard, "Affective Computing," MIT Media Laboratory Perceptual Computing Section, Cambridge, MA, USA, Tech. Rep. 321, Nov. 1995.

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

[3] A. Álvarez, I. Cearreta, J. López, A. Arruti, E. Lazkano, B. Sierra, and N. Garay, "Feature subset selection based on evolutionary algorithms for automatic emotion recognition in spoken spanish and standard basque language," in *Proc. 9th Int. Conf. Text, Speech and Dialogue (TSD 2006)*, Brno, Czech Republic, Sep. 2006, pp. 565–572.

[4] D. Ververidis and C. Kotropoulos, "Fast sequential floating forward selection applied to emotional speech features estimated on DES and SUSAS data collections," in *Proc. XIV Eur. Signal Process. Conf. (EUSIPCO'06)*, Florence, Italy, Sep. 2006, pp. 929–932.

[5] M. Sedaaghi, C. Kotropoulos, and D. Ververidis, "Using adaptive genetic algorithms to improve speech emotion recognition," in *Proc. Int. Workshop Multimedia Signal Process. (MMSP'07)*, Chania, Crete, Greece, Oct. 2007, pp. 461–464.

[6] C. Busso, S. Lee, and S. Narayanan, "Using neutral speech models for emotional speech analysis," in *Proc. Interspeech'07—Eurospeech*, Antwerp, Belgium, Aug. 2007, pp. 2225–2228.

[7] E. Douglas-Cowie, L. Devillers, J. Martin, R. Cowie, S. Savvidou, S. Abrilian, and C. Cox, "Multimodal databases of everyday emotion: Facing up to complexity," in *Proc. 9th Eur. Conf. Speech Commun. Technol. (Interspeech'05)*, Lisbon, Portugal, Sep. 2005, pp. 813–816.

[8] P. Juslin and P. Laukka, "Communication of emotions in vocal expression and music performance: Different channels, same code?," *Psychol. Bull.*, vol. 129, no. 5, pp. 770–814, Sep. 2003.

[9] P. Lieberman and S. Michaels, "Some aspects of fundamental frequency and envelope amplitude as related to the emotional content of speech," *J. Acoust. Soc. Amer.*, vol. 34, no. 7, pp. 922–927, Jul. 1962.

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

[11] A. Paeschke, M. Kienast, and W. Sendlmeier, "F0-contours in emotional speech," in *Proc. 14th Int. Conf. Phonetic Sci. (ICPh'99)*, San Francisco, CA, Aug. 1999, pp. 929–932.

[12] A. Paeschke, "Global trend of fundamental frequency in emotional speech," in *Proc. Speech Prosody (SP'04)*, Nara, Japan, Mar. 2004, pp. 671–674.

[13] T. Bänziger and K. Scherer, "The role of intonation in emotional expressions," *Speech Commun.*, vol. 46, no. 3–4, pp. 252–267, Jul. 2005.

[14] M. Bulut and S. Narayanan, "On the robustness of overall F0-only modification effects to the perception of emotions in speech," *J. Acoust. Soc. Amer.*, vol. 123, no. 6, pp. 4547–4558, Jun. 2008.

[15] K. Scherer, D. Ladd, and K. Silverman, "Vocal cues to speaker affect: Testing two models," *J. Acoust. Soc. Amer.*, vol. 76, no. 5, pp. 1346–1356, Nov. 1984.

[16] C. Busso and S. Narayanan, "Interplay between linguistic and affective goals in facial expression during emotional utterances," in *Proc. 7th Int. Seminar Speech Production (ISSP'06)*, Ubatuba-SP, Brazil, Dec. 2006, pp. 549–556.

[17] G. Kochanski, "Prosody beyond fundamental frequency," in *Methods in Empirical Prosody Research*, ser. Language, Context and Cognition Series, D. Lenertová, R. Meyer, S. Pappert, P. Augurzky, I. Mleinek, N. Richter, and J. Schliesser, Eds. Berlin, Germany: Walter de Gruyter & Co., Apr. 2006, pp. 89–122.

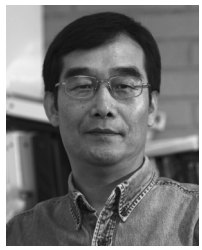
[18] E. Klabbers and J. V. Santen, "Clustering of foot-based pitch contours in expressive speech," in *5th ISCA Speech Synthesis Workshop*, Pittsburgh, PA, Jun 2004, pp. 73–78.

[19] D. Paul and J. Baker, "The design for the Wall Street Journal-based CSR corpus," in *Proc. 2nd Int. Conf. Spoken Lang. Process. (ICSLP'92)*, Banff, AB, Canada, Oct. 1992, pp. 899–902.

[20] J. Garofolo, L. Lamel, W. Fisher, J. Fiscus, D. Pallett, and N. Dahlgren, "TIMIT Acoustic-Phonetic Continuous Speech Corpus,' 1993.
[21] S. Lee, S. Yildirim, A. Kazemzadeh, and S. Narayanan, "An articulatory study of emotional speech production," in *Proc. 9th Eur. Conf. Speech Commun. Technol. (Interspeech'05—Eurospeech)*, Lisbon, Portugal, Sep. 2005, pp. 497–500.
[22] M. Grimm, K. Kroschel, E. Mower, and S. Narayanan, "Primitives-based evaluation and estimation of emotions in speech," *Speech Commun.*, vol. 49, no. 10–11, pp. 787–800, Oct.–Nov. 2007.
[23] M. Liberman, K. Davis, M. Grossman, N. Martey, and J. Bell, "Emotional prosody speech and transcripts," in *Proc. Linguist. Data Consortium*, Philadelphia, PA, 2002, CD-ROM.
[24] F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, and B. Weiss, "A database of German emotional speech," in *9th European Conf. Speech Communication and Technology (Interspeech'2005—Eurospeech)*, Lisbon, Portugal, Sep. 2005, pp. 1517–1520.
[25] J. Montero, J. Gutiérrez-Arriola, S. Palazuelos, E. Enríquez, S. Aguilera, and J. Pardo, "Emotional speech synthesis: From speech database to TTS," in *5th Int. Conf. Spoken Lang. Process. (ICSLP'98)*, Sydney, Australia, Nov.–Dec. 1998, pp. 923–925.
[26] E. Douglas-Cowie, N. Campbell, R. Cowie, and P. Roach, "Emotional speech: Towards a new generation of databases," *Speech Commun.*, vol. 40, no. 1–2, pp. 33–60, Apr. 2003.
[27] C. Busso and S. Narayanan, "Recording audio-visual emotional databases from actors: A closer look," in *Proc. 2nd Int. Workshop Emotion: Corpora for Research on Emotion and Affect, Int. Conf. Language Resources and Evaluation (LREC 2008)*, Marrakech, Morocco, May 2008, pp. 17–22.
[28] P. Boersma and D. Weeninck, "PRAAT, a system for doing phonetics by computer," Inst. Phon. Sci. Univ. of Amsterdam, Amsterdam, Netherlands, Tech. Rep. 132, 1996 [Online]. Available: http://www.praat.org.
[29] 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 Int. Conf. Spoken Lang. Process. (ICSLP'92)*, Banff, AB, Canada, Oct. 1992, pp. 867–870.
[30] S. Ananthakrishnan and S. Narayanan, "Automatic prosody labeling using acoustic, lexical, and syntactic evidence," *IEEE Trans. Speech, Audio, Lang. Process.*, vol. 16, no. 1, pp. 216–228, Jan. 2008.
[31] P. Taylor, "Analysis and synthesis of intonation using the tilt model," *J. Acoust. Soc. Amer.*, vol. 107, no. 3, pp. 1697–1714, Mar. 2000.
[32] E. Grabe, G. Kochanski, and J. Coleman, "Connecting intonation labels to mathematical descriptions of fundamental frequency," *Lang. Speech*, vol. 50, no. 3, pp. 281–310, Oct. 2007.
[33] T. Cover and J. Thomas, *Elements of Information Theory*. New York: Wiley-Interscience, 2006.
[34] R. Duda, P. Hart, and D. Stork, *Pattern Classification*. New York: Wiley-Interscience, 2000.
[35] D. Hosmer and S. Lemeshow, *Applied Logistic Regression*, ser. Probability and Statistics. New York: Wiley, 2000.
[36] N. Nagelkerke, "A note on a general definition of the coefficient of determination," *Biometrika*, vol. 78, no. 3, pp. 691–692, Sep. 1991.
[37] C. Busso and S. Narayanan, "Joint analysis of the emotional fingerprint in the face and speech: A single subject study," in *Proc. Int. Workshop Multimedia Signal Process. (MMSP 2007)*, Chania, Crete, Greece, Oct. 2007, pp. 43–47.
[38] D. Ladd, K. Silverman, F. Tolkmitt, G. Bergmann, and K. Scherer, "Evidence for the independent function of intonation contour type, voice quality, and F0 range in signaling speaker affect," *J. Acoust. Soc. Amer.*, vol. 78, no. 2, pp. 435–444, Aug. 1985.
[39] S. Yildirim, M. Bulut, C. Lee, A. Kazemzadeh, C. Busso, Z. Deng, S. Lee, and S. Narayanan, "An acoustic study of emotions expressed in speech," in *8th Int. Conf. Spoken Lang. Process. (ICSLP'04)*, Jeju Island, Korea, Oct. 2004, pp. 2193–2196.
[40] R. Cowie and R. Cornelius, "Describing the emotional states that are expressed in speech," *Speech Commun.*, vol. 40, no. 1–2, pp. 5–32, Apr. 2003.
[41] M. Grimm, K. Kroschel, and S. Narayanan, "The Vera AM Mittag German audio-visual emotional speech database," in *IEEE Int. Conf. Multimedia Expo (ICME'08)*, Hannover, Germany, Jun. 2008, pp. 865–868.
[42] T. Polzin, "Verbal and non-verbal cues in the communication of emotions," in *Int. Conf. Acoust., Speech, Signal Process. (ICASSP'00)*, Istanbul, Turkey, Jun. 2000, pp. 53–59.

Carlos Busso (S'02–M'09) received the B.S. and M.S. degrees with high honors in electrical engineering from the University of Chile, Santiago, Chile, in 2000 and 2003, respectively, and the Ph.D. degree in electrical engineering from University of Southern California (USC), Los Angeles, in 2008.

Since 2008, he has been a Postdoctoral Research Associate at the Signal Analysis and Interpretation Laboratory (SAIL), USC. He was selected by the School of Engineering of Chile as the best Electrical Engineer graduated in Chile in 2003. At USC, he received a Provost Doctoral Fellowship from 2003 to 2005 and a Fellowship in Digital Scholarship from 2007 to 2008. His research interests are in digital signal processing, speech and video processing, and multimodal interfaces. His current research includes modeling and understanding human communication and interaction, with applications to automated recognition and synthesis to enhance human–machine interfaces. He has worked on audiovisual emotion recognition, analysis of emotional modulation in gestures and speech, designing realistic human-like virtual characters, speech source detection using microphone arrays, speaker localization and identification in an intelligent environment, and sensing human interaction in multiperson meetings.



Sungbok Lee (M'08) received the B.S. degree in chemistry and the M.S. degree in physics from Seoul National University, Seoul, Korea, in 1978 and 1985, respectively, and the Ph.D. degree in biomedical engineering from the University of Alabama, Birmingham, in 1991.

From 1991 to 1997, he was a Research Engineer at the Central Institute for the Deaf, Washington University, St. Louis, MO. From 1998 to 2002, he was with Lucent Bell Labs, Murray Hill, NJ, and with AT&T Labs-Research, Florham Park, NJ, as Research Consultant. Currently, he is a Research Assistant Professor at the Speech Analysis and Interpretation Laboratory (SAIL), Department of Electrical Engineering and a Senior Research Associate at the Phonetics and Phonology Group, Department of Linguistics, University of Southern California, Los Angeles. His research interests include developmental aspects of the speech production, automatic speech recognition, and speech and language processing for human–machine interaction.

Shrikanth Narayanan (S'88–M'95–SM'02–F'09) received the Ph.D. degree from the University of California, Los Angeles, in 1995.

He is Andrew J. Viterbi Professor of Engineering at the University of Southern California (USC), Los Angeles, where he holds appointments as Professor in electrical engineering and jointly in computer science, linguistics, and psychology. Prior to joining USC, he was with AT&T Bell Labs and AT&T Research, first as a Senior Member, and later as a Principal Member of its Technical Staff from 19952000. At USC, he is a member of the Signal and Image Processing Institute and a Research Area Director of the Integrated Media Systems Center, an NSF Engineering Research Center. He has published over 300 papers and has 15 granted/pending U.S. patents.

Dr. Narayanan is a recipient of an NSF CAREER Award, a USC Engineering Junior Research Award, a USC Electrical Engineering Northrop Grumman Research Award, a Provost Fellowship from the USC Center for Interdisciplinary Research, a Mellon Award for Excellence in Mentoring, and a recipient of a 2005 Best Paper Award from the IEEE Signal Processing Society. Papers by his students have won best student paper awards at ICSLP'02, ICASSP'05, MMSP'06, and MMSP'07. He is an Editor for the *Computer Speech and Language Journal* (2007–present) and an Associate Editor for the *IEEE Signal Processing Magazine* and the IEEE TRANSACTIONS ON MULTIMEDIA. He was also an Associate Editor of the IEEE TRANSACTIONS ON SPEECH AND AUDIO PROCESSING (2000–2004). He served on the Speech Processing technical committee (2003–2007) and the Multimedia Signal Processing technical committee (2004–2008) of the IEEE Signal Processing Society and the Speech Communication committee of the Acoustical Society of America (2003–present). He is a Fellow of the Acoustical Society of America and a member of Tau Beta Pi, Phi Kappa Phi, and Eta Kappa Nu.