STATISTICAL ANALYSIS OF EMOTION DETECTION USING FUNDAMENTAL FREQUENCY

A.Vasavi, *Assistant Professor*, J.Leela Mahendra, *Assistant Professor*
and Shaik.Abdul Rahim , *Assistant Professor*
Department of Electronics & Instrumentation Engineering
RGMCET, Nandyal

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.. 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.

Keywords—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).

Speech prosody is one of the important communicative channels that is influenced by and enriched with emotional modulation. The goal of this paper is two fold. The first is to study which aspects of the pitch contour are manipulated during expressive speech (e.g., curvature, contour, shape, and 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 FO contour. 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 prosody speech models to detect emotional speech Gaussian mixture models (GMMs) are trained using the most discriminative aspects of the pitch contour, following the analysis results presented in this paper.

In this paper, we have implemented a method for Statistical analysis of emotion detection using fundamental frequency. The methodology is discussed in section II. In section III, Comparisons using Symmetric KullBack-Leibler distance are described. Section IV covers Logistic Regression Analysis section V covers Emotional Discrimination Results using Neural models section VI concludes the study.

II. METHODOLOGY

A. Overview:

The fundamental frequency or FO contour (pitch), which is a prosodic feature, provides the tonal and rhythmic properties of the speech.

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.

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.

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 artic-ulography (EMA) system. 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 articulator information, only the acoustic signals are analyzed in this study

The second emotional corpus corresponds to the Emotional Prosody Speech and Transcripts database (EPSAT). 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 the neutral state

The third emotional corpus is the Database of German Emotional Speech (GES) which was collected at the Technical University of Berlin . 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 con-sidered in the database are anger, happiness, sadness, boredom, disgust, and fear, in addition to neutral state.

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 (£$_r$ef) ^{an}d the neutral subset in the emotional databases (£^$_{eu}$) 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.

$$S^{s}_{Energy} = \sqrt{\frac{E_{ref}}{E^{s}_{neu}}} \quad \ldots.(1)$$

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 $F0_{re}f$. Then, the average pitch value for the neutral set of the emotional databases is estimated for each speaker $F0_{neu}$. Finally, a scaling factor (Sp_0) is estimated by taking the ratio between $–F0_{ref}$ and $F0_{neu}$.Therefore, the neutral samples of each speaker in the databases will have a similar FO mean value.

$$s_{F0} = F0_{ref} \,/\, F0_{neutral} \ldots(2)$$

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 pre-recorded.

D. Pitch Features:

The pitch contour was extracted with the Praat speech processing software, 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.

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). 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 . Taylor

has proposed an alternative pitch contour parameterization called Tilt Intonation Mode! . 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 limi-tations, we follow a similar approach presented by Grabe et a!. . The voiced regions, which are automatically segmented from the pitch values, are parameterized using polynomials. This parameterization captures the local shape of the FO contour with few parameters, which provides clear physical interpretation of the curves. Here, the slope (tti), curvature, and inflexion (c3) 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

$$y = a1.x + a0 \text{.....(3)}$$

$$y = b2.x^2 + b1.x + b0 \text{.....(4)}$$

$$y = c3.x^3 + c2.x^2 + c1.x + c0 \text{...(5)}$$

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.

III. EXPERIMENT I: 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 . 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.

Since the KLD is not a symmetric metric, we propose the use of the symmetric Kullback-Leibler distance or ^-divergence, which is defined as

$$J_{(q,p)} = D(q//p) + D(p//q)/2 \text{ ..... (6)}$$

Where D(p //q) is the conventional KLD

$$D(q//p) = \sum_{x \in X} q(X) \log(q(x)/p(x)) \text{ .. (7)}$$

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. 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 $J_f^{(d,e)}$ ($p_f^{(d,e)}$, $q_f^{(ref)}$) . This procedure is repeated for each database and for each emotional category.

A good pitch feature for emotion discrimination ideally would have $J_f^{(d,neytral)}$ close to zero (neutral speech of the database d is similar to the reference corpus) and a high value for $J_f^{(d,e)}$,where e is any emotional category except the neutral state. Notice that if $J_f^{(d,neytral)}$ and $J_f^{(d,e)}$, 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 . $J_f^{(d,e)}$,and $J_f^{(d,neytral)}$ .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.

$$r_f^{(d,\bar{e})} = J_f^{(d,\bar{e})} / J_f^{(d,neutral)} \quad \text{.....(8)}$$

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.

IV. EXPERIMENT 2: LOGISTIC REGRESSION ANALYSIS

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 (β)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

$$E(Y / f1, f2 ..... fn) = \pi(x) = \frac{e^{\beta 0+\beta 1x1+....+\beta nxn}}{1+e^{\beta 0+\beta 1x1+....+\beta nxn}} .....(9)$$

$$g(x) = \ln[\pi(x)/1-\pi(x)]$$

$$= ^{\beta 0+\beta 1x1+....+\beta nxn} .....(10)$$

Logistic regression analysis is used witb/or-ward feature selection (FFS) to discriminate between each emotional category and neutral state (i.e., neutral-anger).

V. EMOTIONAL DISCRIMINATION RESULTS USING NEUTRAL MODELS

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 .

A. Motivation and Proponed 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 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 .

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 & 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.

FO 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.

The maximum likelihood estimates of the parameters in the GMM 8 are computed using the expectation-maximization (EM) algorithm. For a given input speech, the likelihoods of the models, Ff(Xt — x/|©), are used as fitness measures. In the second step, a Linear Discriminate Classifier (LDC) was implemented to discriminate between neutral and expressive speech. For a given input speech, the likelihoods of the models, Ff(Xt — x/|©), are used as fitness measures. In the second step, a Linear Discriminate Classifier (LDC) was implemented to discriminate between neutral and expressive speech.

$$F_f(X_f = x_f|\Theta) = \sum_{j=1}^{k} \alpha_j \frac{1}{\sigma_j\sqrt{2\pi}} \exp(\frac{(X_f-\mu_j)^2}{-2\sigma_j^2}) ...(12)$$

with

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

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 An important parameter of the GMM is the number of mixtures, 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.

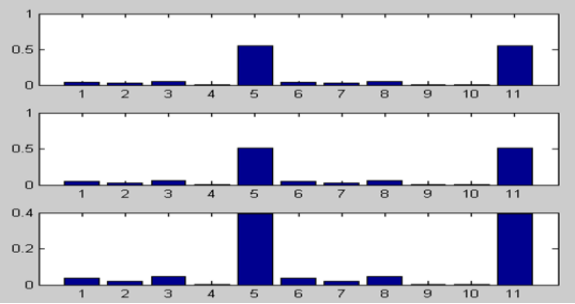

Fig. 1 Statistical analysis for neutral and emotional Speech

TABLE I

CONFUSION MATRIX FROM SUBJECTIVE HUMAN EVALUATION [23].

Stimuli	Correctly classified responses (%)				
	Anger	Happ.	Neutral	Sadness	Surprise
Anger	**75.1**	4.5	10.2	1.7	8.5
Happiness	3.8	**56.4**	8.3	1.7	29.8
Neutral	4.8	0.1	**60.8**	31.7	2.6
Sadness	0.3	0.1	12.6	**85.2**	1.8
Surprise	1.3	28.7	10.0	1.0	**59.1**
Total rate	67.3%				

TABLE II

CONFUSION MATRIX FOR THE BAYES CLASSIFIER WITH SFFS WHEN CROSS-VALIDATION REPETITIONS ARE LIMITED TO 30 AND 30% OF THE UTTERANCES ARE USED FOR TESTING [9].

Stimuli	Correctly classified responses (%)				
	Anger	Happ.	Neutral	Sadness	Surprise
Anger	**41.65**	19.28	16.20	11.05	11.82
Happiness	19.24	**32.19**	18.29	11.04	19.24
Neutral	7.28	5.88	**47.63**	31.09	8.12
Sadness	2.03	1.52	18.32	**72.79**	5.34
Surprise	22.28	14.40	7.33	14.94	**41.05**
Total rate	47.06%				

TABLE III

CONFUSION MATRIX WHEN THE ADAPTIVE GA REMOVES THE MEAN VALUE OF THE SECOND FORMANT AND UTTERANCES 1132-1135 FROM SUBSEQUENT CLASSIFICATION.

Stimuli	Correctly classified responses (%)				
	Anger	Happ.	Neutral	Sadness	Surprise
Anger	**44.40**	17.43	14.39	10.09	13.69
Happiness	18.86	**37.73**	11.79	12.34	19.28
Neutral	4.79	5.75	**47.81**	36.17	5.48
Sadness	2.40	2.40	19.63	**71.57**	4.00
Surprise	14.62	18.90	10.76	12.69	**43.03**
Total rate	48.91%				

VI. Conclusion

This paper presented an analysis of different expressive pitch contour statistics with the goal of finding the emotionally salient aspects of the FO 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.

higher than the neutral samples.

REFERENCES

[1] J 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. Fellcnz, and J. Taylor, "Emoiion recognition in human-computer interaction." IEEE Signal Process. Mag., vol. 18, no. 1- pp. 32-80, Jan. 2001.

[3] A. Alvarez, I. Cearreta, J. Lopez, A. Arruti, E. Lazkano, B. Sierra, and N. Garay, "Feature subset selection based on evolutionary algorithm., 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. Vervcridis and C. Kotropoulos, "Fast sequential floating forward selection applied lo emotional .speech features estimated on DES and S US AS data collect ions," in Prot.. XIV Ear. Signal Process. Conf. (EU-SiPCO'Of,), 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