IOSR Journal of Electrical and Electronics Engineering (IOSR-JEEE)
e-ISSN: 2278-1676, p-ISSN: 2320-3331
PP 36-40
www.iosrjournals.org

Salient Features Extraction for Emotion Detection Using Modified Kull Back Leibler Divergence

MD TOUSEEF SUMER[1]
[1] *Dept of ECE, JNTUH, India*

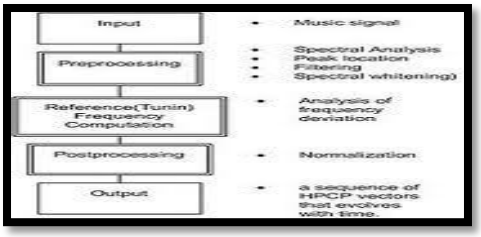
ABSTRACT : *During a speech voice is enriched to convey not only intended language message but also the emotion state of a person. Recent advancements suggested that emotion is a integral to our rational and intelligent decision, which helps to relate expressing our feelings. Speech prosody is one of the important communicative channel influences to express the emotional message. Pitch contour is one of the important features of speech that is affected by modulation by which emotions can be recognized whether it may be a neutral or emotion speech. In this project the pitch contour is analyzed in order to extract salient features for emotion detection. Fundamental frequency F_0 is plotted based upon pitch contour features like mean, minimum, maximum, standard deviation, kurtosis, upper quartile ,lower quartile, skewness and inflexion. Here F_0 is a rhythmic property of speech source using global statistics of pitch contour over entire utterance or sentence known as Sentence Level. Entrance of pitch features over voiced region is known as Voiced Level. Probability density function has to be plotted for each feature and for all emotional and reference data bases. In order to discriminate neutral and emotional speech KLD is used. If the distance between reference neutral and emotion are similar it is best features. If the distance between reference neutral and emotion are different then it is lowest.GMM which is a parametric model used for building these best features and testing the GMM to identify the emotional speech. Result will be indicating that emotional modulation is not uniformly distributed in time and space across different communicative languages.*

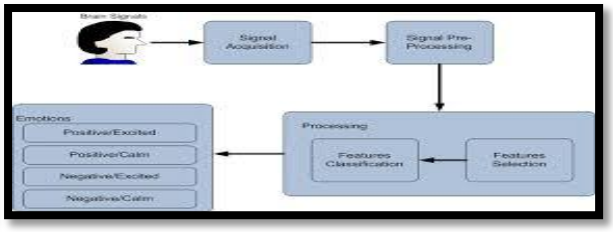
Keywords: *Emotional speech analysis, emotional speech recognition, expressive speech, KLD, pitch contour analysis.*

I. INTRODUCTION

Emotion Speech not only contains the literal meaning (semantics) of the words spoken, but convey a wealth of additional information to the listener. Emotions colour the language, and can make meaning more complex. As listeners we also react to the speakers emotive state and adapt our behavior depending on what kind of emotions the speaker transmit, e.g. we may try to show empathy to sad people, or if someone hesitates we try to make the person clarify what she/he means or wants. To classify the emotive state by a speaker on basis of the prosody and voice quality, we have to classify acoustic features in the speech as connected to certain emotions. This also implies the assumption that voice alone really carries full information about emotive state by the speaker. The human listener is particularly finely tuned to the detection of a range of emotions in the speaker's utterance. A native listener is able to detect certain emotions by recognizing salient words, which are associated with those emotions. 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 intonation, tone, timing, and energy of speech are all jointly influenced in a nontrivial manner to express the emotional message. In this paper we analyze the importance of different acoustic and prosodic measurements, i.e. we examine expressive patterns that are based on vocal intonation. This paper focuses on one aspect of expressive speech prosody, the F0 (pitch) contour. The first is to analyze which aspects of the pitch contour are interpolated during expressive speech (e.g., curvature, contour, shape, and dynamics). For this purpose, we present a *Kullback–Leibler divergence* (KLD) .to 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. Modified Kullback–Leibler distance is used to compare the distributions of different pitch statistics (e.g., mean, maximum) between emotional speech and reference neutral speech. To build robust emotionally salient features we use Gaussian mixture models. Gaussian mixture models (GMMs) are trained using the most discriminative aspects of the pitch contour, 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 pitch shape itself (e.g., slope, curvature, and inflexion). 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

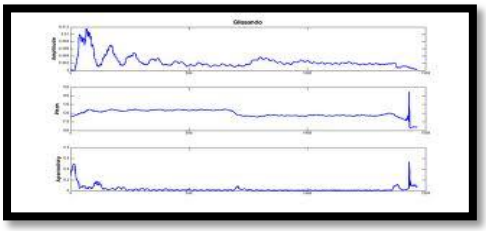
than when they are trained with features estimated over shorter speech regions. Using the most salient pitch features, the performance of the proposed approach for binary emotion recognition reaches over 90% (baseline 77%), when the various acted emotional databases are considered together.

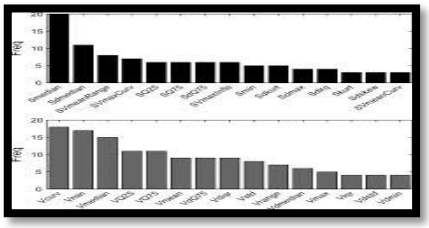




II. FUNDAMENTAL FREQUENCY OR F0 CONTOUR (PITCH)

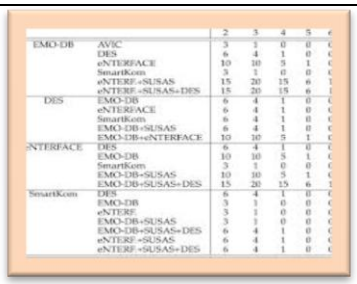
The fundamental frequency or F0 contour (pitch), which is a prosodic feature, provides the tonal and rhythmic properties of the speech. It predominantly describes the speech source rather than the vocal tract properties. Although it is also used to emphasize linguistic goals conveyed in speech, it is largely independent of the specific lexical content of what is spoken in most languages. Fundamental frequency F_0 is plotted based upon pitch contour features like mean, minimum, maximum, standard deviation, kurtosis, upper quartile ,lower quartile, skewness and inflexion. Here F_0 is a rhythmic property of speech source using global statistics of pitch contour over entire utterance or sentence known as *Sentence Level*. Entrance of pitch features over voiced region is known as *Voiced Level*. Probability density function has to be plotted for each feature and for all emotional and reference data bases In this approach, the frames are labeled as voiced or unvoiced frames according to their F0 value (greater or equal to zero). 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 signals.



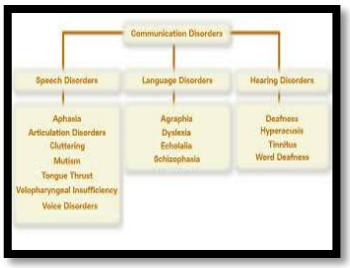


III. DATABASES

Databases are existing works on anger recognition one has to be aware of essential conditions in underlying database design. The most restricted database settings would certainly have prearranged sentences performed by professional speakers (one at a time) recorded in audio studio.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 I the pitch. The first database was collected using an *electromagnetic Articulography* (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 articulatory information, only the acoustic signals are analyzed in this study.



		2	3	4	5	6
EMO-DB	AVIC	3	1	0	0	0
	DES	6	4	1	0	0
	eNTERFACE	10	10	5	1	0
	SmartKom	3	1	0	0	0
	eNTERF+SUSAS	15	20	15	6	1
	eNTERF+SUSAS+DES	15	20	15	6	1
DES	EMO-DB	6	4	1	0	0
	eNTERFACE	6	4	1	0	0
	SmartKom	6	4	1	0	0
	EMO-DB+SUSAS	6	4	1	0	0
	EMO-DB+eNTERFACE	10	10	5	1	0
NTERFACE	DES	6	4	1	0	0
	EMO-DB	10	10	5	1	0
	SmartKom	3	1	0	0	0
	EMO-DB+SUSAS	10	10	5	1	0
	EMO-DB+SUSAS+DES	15	20	15	6	1
SmartKom	DES	6	4	1	0	0
	EMO-DB	3	1	0	0	0
	eNTERF	3	1	0	0	0
	EMO-DB+SUSAS	3	1	0	0	0
	EMO-DB+SUSAS+DES	6	4	1	0	0
	eNTERF+SUSAS	6	4	1	0	0
	eNTERF+SUSAS+DES	6	4	1	0	0



IV. 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.

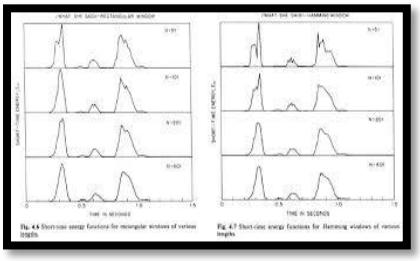
In the first step, the speech files are scaled such that the averageRMS energy of the neutral reference database and the neutral subset in the emotional databases are the same for each speaker . This normalization is separately applied for each subject in each database. The goal of this normalization is to compensate for different recording settings among the databases.

$$S^s_{\text{Energy}} = \sqrt{\frac{E_{\text{ref}}}{E^s_{\text{neu}}}}. \qquad S^s_{F0} = \frac{F0_{\text{ref}}}{F0^s_{\text{neu}}}.$$

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.

V. PITCH FEATURES

The pitch contour was extracted with the Praat speech processing software using an autocorrelation method. Pitch is fundamental frequency of speech signal. The most widely considered areas of stress evaluation consider the characteristics of pitch These studies consider subjective assessment of pitch frequency, statistical analysis of pitch mean, variance, and distribution The pitch signal depends on the tension of the vocal folds and the sub glottal air pressure when speech is generated. The pitch signal is produced due to the vibration of the vocal folds. Hence by using above we can calculate pitch features easily. Formants are defined as the spectral peaks of sound spectrum, of the voice, of a person. In speech science and phonetics, formants frequencies refer to the acoustic resonance of the human vocal tract .They are often measured as amplitude peaks in the frequency spectrum of the sound wave. We have considered the first 3 formants f1, f2, f3 for analysis of emotions. For different vowels, the range of f1 lies between 270 to 730 Hz while the range of f2 and f3 lie between 840 to 2290 and 1690 to 3010 Hz respectively[.Formant frequencies are very much important in the analysis of the emotional state of a person. The Linear predictive coding technique (LPC) has been used for estimation of the formant frequencies.



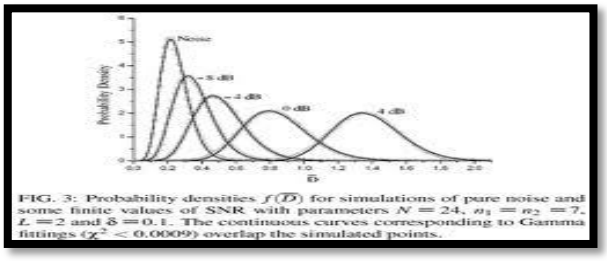


FIG. 3: Probability densities $f(\overline{D})$ for simulations of pure noise and some finite values of SNR with parameters $N = 24$, $m_1 = m_2 = 7$, $\mathcal{L} = 2$ and $\mathcal{S} = 0.1$. The continuous curves corresponding to Gamma fittings ($\chi^2 < 0.0009$) overlap the simulated points.

IOSR Journal of Electrical and Electronics Engineering (IOSR-JEEE)
e-ISSN: 2278-1676, p-ISSN: 2320-3331
PP 36-40
www.iosrjournals.org

VI .KULLBACK–LEIBLER DISTANCE

The Kullback–Leibler divergence (or information divergence, or information gain, or relative entropy) is a natural distance measure from a ”true” probability distribution p to an arbitrary probability distribution q. Typically p represents data, observations, or a precise calculated probability distribution. The measure q typically represents a theory, a model, a description or an approximation of p. originally introduced by Solomon Kullback and Richard Leibler in 1951, as the directed divergence between two distributions, 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.

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

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

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

A good pitch feature for emotion discrimination ideally would have close to zero (neutral speech of the Database is similar to the reference corpus) and a high value. The pitch features with higher values are *SV mean Min*, *SV mean Max*, *Sdiqr*, and *Smean* for the sentence- level features and *Vrange*, *Vstd*, *Vdrange*, and *Vdiqr* for the voiced-level features.

VII. GAUSSIAN MIXTURE MODELS

A Gaussian Mixture Model (GMM) is a parametric probability density function represented as a weighted sum of Gaussian component densities. GMMs are commonly used as a parametric model of the probability distribution of continuous measurements or features in a biometric system, such as vocal-tract related spectral features in a speaker recognition system. The GMM parameters are iteratively estimated using the expectation maximization (EM) algorithm. The posterior probabilities of each bin being generated by the target mixtures are estimated, and then these values are used to calculate the Gaussian component parameters. These steps are iterated to converge upon a solution.

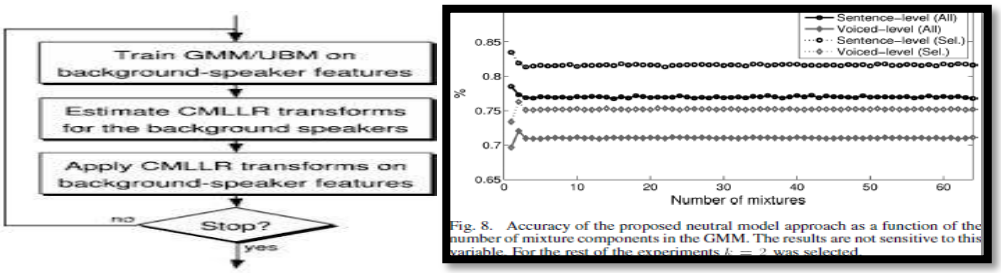


Fig. 8. Accuracy of the proposed neutral model approach as a function of the number of mixture components in the GMM. The results are not sensitive to this variable. For the rest of the experiments $\lambda = 2$ was selected.

VIII. RESULTS

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 78% 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. 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). We hypothesized that neutral speech prosodic models trained with English speech can be used to detect emotional speech in another language.

IX. 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, experiments were proposed. In the experiment, the distribution of different pitch features was compared with the distribution of the features derived from neutral speech using the symmetric KLD. The 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. Results from our previous work indicated that emotional modulation is not uniformly distributed, in time and in space, across different communicative channels. If this trend is also observed in the fundamental frequency, certain regions in the pitch contour might present stronger emotional modulation .And we get result up to 85%.

REFERENCES

[1] J. Garofolo, L. Lamel, W. Fisher, J. Fiscus, D. Pallet, and N. Dahlgren, "TIMIT Acoustic- Phonetic Continuous Speech Corpus,' 1993.
[2] 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.
[3] 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.
[4] M. Liberian, K. Davis, M. Grossman, N. Marty, and J. Bell, "Emotional prosody speech and transcripts," in *Proc. Linguist. Data Consortium*,
Philadelphia, PA, 2002, CD-ROM.
[5] F. Burckhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, and B. Weiss, "A data base of German emotional speech," in *9th European Conf. Speech Communication and Technology (Interspeech'2005—Eurospeech)*, Lisbon, Portugal, Sep. 2005, pp. 1517–1520.
[6] J. Montero, J. Gutiérrez-Arriola, S. Palazuelos, E. Enrique, S. Aguilera, and J. Pardon, "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.
[7] 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.
[8] S. Ananthakrishnan and S. Narayanan, "Automatic prosody labeling Using acoustic, lexical, and syntactic evidence," *IEEE Trans. Speech*.