Cross-Lingual Vocal Emotion Recognition in Five Native Languages of Assam Using Eigenvalue Decomposition

Aditya Bihar Kandali[1], Aurobinda Routray[1], and Tapan Kumar Basu[2]

[1] Department of Electrical Engineering, Indian Institute of Technology Kharagpur, PIN Code-721302, India
abkandali@rediffmail.com, aroutray@ee.iitkgp.ac.in
[2] Aliah University, Salt Lake City, Kolkota, India
basutk02@yahoo.co.in

Abstract. This work investigates whether vocal emotion expressions of full-blown discrete emotions can be recognized cross-lingually. This study will enable us to get more information regarding nature and function of emotion. Furthermore, this work will help in developing a generalized vocal emotion recognition system, which will increase the efficiency required for human-machine interaction systems. An emotional speech database was created with 140 simulated utterances (20 per emotion) per speaker, consisting of short sentences of six full-blown discrete basic emotions and one 'no-emotion' (i.e. neutral) in five native languages (not dialects) of Assam. A new feature set is proposed based on Eigenvalues of Autocorrelation Matrix (EVAM) of each frame of utterance. The Gaussian Mixture Model is used as classifier. The performance of EVAM feature set is compared at two sampling frequencies (44.1 kHz and 8.1 kHz) and with additive white noise with signal-to-noise ratios of 0 db, 5 db, 10 db and 20 db.

Keywords: Full-blown Basic Emotion, Cross-lingual Vocal Emotion Recognition, Gaussian Mixture Model, Eigenvalues of Autocorrelation Matrix.

1 Introduction

Human beings express emotions explicitly in speech, face, gait and other body languages. The vocal expressions are harder to regulate than other explicit emotional signals. So, it is possible to know the actual affective state of the speaker from her/his voice without any physical contact. But exact identification of emotion from voice is very difficult due to several factors. The speech consists broadly of two components coded simultaneously: (i) "What is said" and (ii) "How it is said". The first component consists of the linguistic information pronounced as per the sounds of the language. The second component consists of non-linguistic or paralinguistic or supra-segmental component which includes the prosody of the language i.e. pitch, intensity and speaking-rate rules to give lexical and grammatical emphasis for the spoken messages; and the prosody of emotion to express

S. Chaudhury et al. (Eds.): PReMI 2009, LNCS 5909, pp. 519–524, 2009.

the affective state of the speaker. In addition, speakers also possess their own style, i.e. a characteristic articulation rate, intonation habit and loudness characteristic. The voice contains also information about the speaker's identity, age, gender, and body size.

The present work investigates a specific research question concerning vocal emotion recognition: "Are vocal emotion expressions of discrete emotions recognized cross-lingually?". This study will enable one to get more information about the nature and function of emotion. It will also help in developing a generalized voice emotion recognition system, which will increase the efficiency of human-machine interaction systems. Some applications are as follows: (i) to obtain more efficient and more accurate performance of automatic speech recognition and automatic speaker recognition systems, due to reduction of search space to models corresponding to pre-recognized emotions [1, 2]; (ii) to design an automatic speech translator across languages retaining the emotional content [1], (iii) to make more efficient automatic tutoring, alerting, and entertainment systems [3]. Picard [4] has explained about affective computing algorithms which can improve problem solving capability of a computer and make it more intelligent by giving it the ability to recognize and express emotions.

When a machine is trained with emotion utterances of one set of languages and tested with emotion utterances of a set of different languages, the process is called as cross-lingual (or cross-cultural) vocal emotion recognition. Very few studies of cross-lingual (i.e. cross-cultural) voice emotion recognition have been reported by researchers [5]. Among these noteworthy is the study by Scherer et al. [6], conducted in nine countries in Europe, the United States, and Asia on vocal emotion portrayals of anger, sadness, fear, joy, and neutral voice, which are produced by professional German actors. In this study, overall perception accuracy by human subjects is found to be 66%. Also, the patterns of confusion are found very similar across all countries, which suggest the existence of similar inference rules from vocal expression across cultures. Generally, accuracy decreases with increasing language dissimilarity from German in spite of the use of language-free utterances. So their conclusion is that culture- and language-specific paralinguistic patterns may influence the decoding process. Juslin and Laukka [5] also reported that cross-cultural decoding accuracy of voice expression of emotions is significantly higher than that expected by chance. Laukka [7] has reported that: (i) vocal expressions of discrete emotions are universally recognized, (ii) distinct patterns of voice cues correspond to discrete emotions, and (iii) vocal expressions are perceived as discrete emotion categories but not as broad emotion dimensions. All the above experiments are done mostly with a very few number of European and Asian languages. So, these findings need to be verified using more number of languages.

The present study is based on a modified Brunswikian lens model of process of vocal communication of emotion [8]. This model motivates research to determine the proximal cues i.e. the representation of voice acoustic cues in the basilar membrane of the cochlea, amygdala, and auditory cortex, which will lead to the perception of the vocal emotion. Based on studies by researchers [3, 8, 9], one

can identify three broad types of proximal voice cues: (i) fundamental frequency or pitch frequency (F0) contour, (ii) continuous acoustic variables: magnitude of fundamental frequency, intensity, speaking rate, and spectral energy distribution; and (iii) voice quality (tense, harsh or breathy): described by high frequency energy, formant frequencies, precision articulation and glottal waveform. A description of relationships among archetypal emotions and the voice cues is given in [3, 8, 9].

In this paper, a new feature set is proposed based on 5 most significant Eigenvalues of Autocorrelation Matrix (EVAM) of each frame of utterance for automatic vocal emotion recognition. The 5 most significant EVAM of a signal represent the powers of 5 most prominent frequency components (though with some additive noise) in the signal [10]. The source-filter model of speech production describes speech as an acoustic excitation signal filtered due to resonances of the vocal tract. The vocal tract resonances are called formant frequencies, or formants; which are the prominent frequencies having relatively higher amplitudes than other frequency components in the speech signal. In general, a speech signal contains 5 to 6 formants. Hence, 5 or 6 most significant EVAM of a short-time frame of speech will represent the powers corresponding to the formants (though with some additive noise), if they are present in that speech frame. The EVAM feature set is also expected to be robust in presence of noise, since these eigenvalues corresponds to the most prominent signal subspace eigenvectors.

The Gaussian mixture model (GMM) classifier is used for classification [11]. The study of cross-lingual vocal emotion recognition is carried out using simulated utterances of 6 full-blown discrete basic emotions (*anger*, *disgust*, *fear*, *happiness*, *sadness* and *surprise*) and 1 '*no−emotion*' (i.e. *neutral*) in 5 Indian languages: Assamese, Bodo (or Boro), Dimasa, Karbi, and Mishing (or Mising), which are the native languages (not dialects) of the state of Assam. The performance of the EVAM feature set is compared at two sampling frequencies (44.1 kHz and 8.1 kHz) and with additive white noise with signal-to-noise ratios of 0 db, 5 db, 10 db and 20 db.

2 Data Collection

As a part of this research work, emotional utterances in native languages of Assam are collected as described below. The subjects are chosen mostly from students and faculty members of some educational institutions of Assam. Some subjects are lay actors and others are trained for the first time through a few rehearsals, so as to avoid exaggerated expressions. Thirty randomly selected volunteer subjects (3 males and 3 females per language) are requested for recording emotional utterances of the 5 native languages of Assam. The utterances are recorded in an almost noise-free small closed room with headphone-mic and notebook computer in a single channel with 44.1 kHz sampling frequency and 16 bit depth. Each subject is asked to utter a fixed set of 140 short sentences (20 per emotion) of variable length of her/his first language only. The subjects are asked to rehearse their acting a few times before final recording.

3 Listening Test

A listening test of the emotional speeches is carried out with the help of 6 randomly selected volunteer listeners (3 Males and 3 Females) for each language of the Multilingual ESDNEI database. The listeners have never heard the speeches of the languages of the Multilingual ESDNEI database. Some of the listeners are selected as different persons for each language while others remained the same, because of the unavailability of a complete set of different volunteer listeners for each language, who do not understand or never heard speeches in the above languages. The average scores of the listening test are given in Table 1.

Table 1. Percentage Cross-lingual Average Recognition Success Scores of Listening Test of the Utterances of the individual languages of the Multilingual ESDNEI database by 6 Human subjects (3 Males and 3 Females) who never heard any of these languages

Language→ Emotion↓	Assamese	Bodo	Dimasa	Karbi	Mishing	Average
Anger	97.22	70.97	95.42	88.19	86.67	87.69
Disgust	95.00	45.83	85.97	75.69	75.42	75.58
Fear	97.78	85.28	95.56	87.08	93.33	91.81
Happiness	85.69	74.86	84.17	78.75	79.44	80.58
Sadness	97.50	88.19	97.08	96.25	94.58	94.72
Surprise	87.78	67.36	78.06	69.58	71.81	74.92
Neutral	99.03	81.81	94.17	93.61	90.97	91.92
Average	94.29	73.47	90.06	84.17	84.60	85.32

4 Experiment: Cross-Lingual Vocal Emotion Recognition

A total of seven GMMs, one for each emotion, are trained using the Expectation-Maximization (EM) algorithm [11] and Leave-One-Out (LOO) cross-validation method [12], and EVAM feature vectors of 10 utterances of the subjects of 4 languages. After training, the classifier is tested with EVAM feature vectors of test utterances consisting of other 10 utterances of the subjects of the left-out language (test-language) as follows. The mean-log-likelihood of EVAM feature vectors of one test-utterance with respect to the trained GMM corresponding to each emotion-class is computed. The test-utterance is considered to belong to that emotion-class with respect to which the mean log-likelihood becomes the largest. The Percentage Average Recognition Success Score (PARSS) of each emotion and the Mean-PARSS (MPARSS) of all emotions are computed from the Recognition Success Scores (RSS) for all 5 combinations of train-test data. The initial means and the elements of diagonal covariance matrices of the GMM are computed by split-Vector Quantization algorithm [13]. The above procedure is repeated for GMMs with different number of components of Gaussian probability distribution functions i.e. M=8, 16 and 32, and the best result is considered. Henceafter, the Mean-PARSS will be referred as the 'Average'.

5 Feature Extraction

In this paper, the speech is preprocessed by detecting the end points and removing the silence periods and the dc component. The frame duration is chosen as 23.22 ms in case of sampling frequency of 44.1 kHz. The utterances are decimated to sampling frequency 8.1 kHz and in this case the frame duration is chosen as 31.6 ms. All the frames are rectangular windowed. The frames are taken with 50% overlaps with neighboring frames. For each frame, the autocorrelation matrix with lag p=32 in case of sampling frequency of 44.1 kHz and lag p=8 in case of sampling frequency of 8.1 kHz are computed. Then after eigen decomposition of the autocorrelation matrix, a 5-element feature vector is formed using the 5 most significant eigenvalues, from each frame. The EVAM features are normalized by subtracting mean and dividing by the standard deviation.

6 Results and Discussion

The percentage average scores of cross-lingual voice emotion recognition in each case are shown in Table 2. It can be observed that the performance of the proposed feature set for the case of original utterances at 8.1 kHz sampling frequency is a little better than at 44.1 kHz sampling frequency. The performance gradually decreases as the Signal-to-Noise Ratio (SNR) is reduced from 20 db to 0 db. It is observed that the performance is satisfactorily above that of the average human recognition score (i.e. 85.32% in Table 1) in the listening test. The results show high potential of EVAM features for emotion recognition from telephone channel voices which have 8 kHz sampling frequency.

Table 2. Percentage Average Score of Cross-lingual Vocal Emotion Recognition from original speech with Noise-not-Added (NNA) and in presence of additive white noise of 4 Signal-to-Noise Ratios (SNRs) [fs: Sampling Frequency, Number of Components in GMM: 32]

SNR (db)→	NNA	NNA	20	10	5	0
fs (kHz)→ Emotion↓	44.1	8.1	8.1	8.1	8.1	8.1
Anger	100.00	100.00	100.00	97.33	95.33	84.33
Disgust	99.67	99.00	97.33	91.67	83.67	74.33
Fear	99.67	100.00	100.00	99.33	98.33	94.33
Happiness	100.00	100.00	100.00	99.67	98.00	96.00
Sadness	98.67	99.67	99.33	98.00	97.00	87.00
Surprise	100.00	100.00	99.33	95.33	91.00	85.67
Neutral	99.00	99.00	98.33	97.67	95.67	93.67
Average	99.57	99.67	99.19	97.00	94.14	87.90

7 Conclusion

This study verified that the full-blown discrete basic vocal emotions are recognized cross-lingually with accuracies much above the chance level. It is also verified that there exist distinct patterns of voice cues corresponding to full-blown discrete basic emotions. The EVAM features have high potential for vocal emotion recognition in telephone channel.

Acknowledgment

The authors are grateful to the students and the faculty members of Jorhat Engineering College, Jorhat, Assam, and all the other people of Assam, who have actively cooperated during the collection of data for the present work. The authors also express gratefulness to all the students of Indian Institute of Technology Kharagpur, India, who have voluntarily taken part as listeners in the listening test.

References

[1] Holmes, J., Holmes, W.: Speech Synthesis and Recognition, 2nd edn. Taylor & Francis, New York (2001)
[2] Rose, P.: Forensic Speaker Identification, p. 302. Taylor & Francis, New York (2002)
[3] Cowie, R., Douglas-Cowie, E., Tsapatsoulis, N., Votsis, G., Kollias, S., Fellenz, W., Taylor, J.G.: Emotion recognition in human-computer interaction. IEEE Signal Process. Mag. 18(1), 32–80 (2001)
[4] Picard, R.W.: Affective Computing. The MIT Press, Cambridge (1997)
[5] Juslin, P.N., Laukka, P.: Communication of Emotions in Vocal Expression and Music Performance. Psychological Bulletin 129(5), 770–814 (2003)
[6] Scherer, K.R., Banse, R., Wallbott, H.G.: Emotion Inferences from Vocal Expression Correlate Across Languages and Cultures. J. Cross-Cultural Psychology 32(1), 76–92 (2001)
[7] Laukka, P.: Vocal Expression of Emotion – Discrete-emotion and Dimensional Accounts. Comprehensive Summaries of Uppsala Dissertations from the Faculty of Social Sciences 141, ACTA Universitatis Upsaliensis, Uppsala (2004)
[8] Scherer, K.R., Johnstone, T., Klasmeyer, G.: Vocal Expression of Emotion. In: Davidson, R.J., Scherer, K.R., Goldsmith, H.H. (eds.) Handbook of Affective Science, Part IV, ch. 23, 1st edn. Oxford University Press, Oxford (2003)
[9] Ekman, P.: Basic Emotions. In: Dalgleish, T., Power, M. (eds.) Handbook of Cognition and Emotion, ch. 3. John Wiley & Sons, Ltd., Sussex (1999)
[10] Marple Jr., S.L.: Digital Spectral Analysis With Applications. Prentice Hall Inc., Englewood Cliffs (1987)
[11] Reynolds, D.A., Rose, R.C.: Robust Text-Independent Speaker Identification Using Gaussian Mixture Speaker Models. IEEE Trans. Speech Audio Process. 3(1), 72–83 (1995)
[12] Fukunaga, K.: Introduction to Statistical Pattern Recognition, 2nd edn. Morgan Kaufmann, Academic Press, New York (1990)
[13] Linde, Y., Buzo, A., Gray, R.M.: An Algorithm for Vector Quantizer Design. IEEE Transactions on Communications 28(1), 84–95 (1980)