Journal Of Harmonized Research (JOHR)

Journal Of Harmonized Research in Engineering
2(4), 2014, 372-376

ISSN 2347 – 7393

Original Research Article

A MODEL & DESIGN OF A DATABASE OF FUNCTIONAL AND EMOTIONAL INTONATION WITH REFERENCE TO ASSAMESE LANAGUAGE

Bimal Kumar Kalita, Laba kr. Thakuria, Barnali Kalita, Purnendu Acharjee, P.H.Talukdar

Department of Instrumentation and USIC
Gauhati University, India

Abstract: - In the speech processing research the modeling and the design of a speech database has an important and necessary role. Through this we are going to give a model to design anddevelop a database of functional and emotionalintonation in Assamese language. Here we have considered the database on conversations from TV and Movies about10 hours and the Utterances of speech are segmented. After speech are segmented the syllable andprosody labeling is evaluated. Then we have considered the functions like statements orquestions and emotions like happiness or angeretc. Our databasewill be applicable forfunctional and emotional intonation recognition.

Keywords— assamese, database, emotion, intonation,

I. Introduction

Intonation has a great significance role in the field of linguistic and speech research technology, the comprehensive study to the pitch of continuous speech is useful to both phonetic study and to the improvement of the naturalness of speech synthesis and it is also important in improving the correctness of speech recognition. In order to improve the naturalness of synthesized speech is a major task in nowadays in speech engineering project. The fundamental task for it is to have a clear understanding on the pitch pattern of natural speech. A speech database is an important resource for speech processing [1]. In recent era the corpus development has been boomed for many languages such as English, Assamese, and Bodo. In Bodo, the research and development of information processing has made considerable progress, and the key to the development lies in building and sharing resource. Speech corpora construction is of great importance in language information processing, speech synthesis and speech recognition. To list some, there are, a spontaneous conversation corpus, a Assamese speech corpus with four regional accents, an undergraduate Mandarin speech database for speaker recognition research, and a Assamese conversational corpus for spontaneous speech recognition, etc. However, there is still no corpora specialized for research on Assamese

For Correspondence:
thakurialabaATgmail.com
Received on: October 2014
Accepted after revision: November 2014

functional and emotional into nation. In this paper, the development of a Database of Functional and Emotional Intonation in Assamese is described, including data collection and labeling.

II. Database Development

Normally there are two kinds of speech database i.e. read speech database and spontaneous speech database. In the first, speakers are asked to read out the specified texts, so the naturalness of the voice is not high and for the second, the voice of the speakers could be more natural. To improve the naturalness of synthesized speech is a major task in nowadays speech engineering project. Telephone conversations or free talks among friends are natural, but it is hard to have a great variety of emotional utterances in these kinds of conversations or talks but the collection of emotional speech is a great problem in corpus construction. An important point is that 85% of all the data in the corpus has tagged as "neutral", which indicates the difficulty of collecting spontaneous and diverse emotional speech. There are various emotional conversations in movies and TV. Therefore, Assamese corpus is developed on the basis of movie and TV play conversations.

A) Video Collection

There are various Assamese video available in video shops, but for our project, there are some criteria in picking suitable video. Firstly, only those in standard Assamese, i.e. in Mandarin, are used. As Assamese corpus is focused on the intonation of standard Assamese .Secondly, only those on modern life themes are used, with those on historical stories rejected. In the historical movies and TV plays, some of the characters. Thirdly, some of the serial TV plays include dozens of episodes. In this case, only one episode is used. There should be a great variety of themes to ensure that all kinds of functions, emotions and speaking styles are covered. Besides this, the intonation database should include different characters, like workers, assistants, engineers, clerks, taxi drivers, teachers, etc. to ensure the variability of speaking styles. The relationship between the characters may also affect the speaking styles. For example, there are relations like boss and employee, friends, father and son, doctor and patients. What is more, people tend to speak differently at different locations. For example, the speaking style at a company meeting may be different from that at home. Therefore, in video collecting, there should be a great variety of themes. The number and length of collected movie and TV plays for Assamese corpus are shown in Table1.

Theme	Movie	TV	Length
Action	7	7	14
Affection	8	7	15
Family	6	7	13
City life	3	12	15
Biography	1	9	10
Comedy	7	3	10
War	3	8	11
Legend	10	12	22
Total	45	65	110

Table 1: Number of Movie & TV play

B) Format Transfer

At the time of collection of videos, they are transferred to audio format. In this research, the audio is saved as wav file, mono, with a sampling rate of 4.41 KHz, in an other folder. The audio files keep the same file names as the video, with different extension, and their lengths are exactly the same. In case of a movie, sometimes the characters and scenes change quickly, and it is hard to detect this only by listening to the audio. In this case, the video may help us know the right speaker of an utterance and the exact situation.

C) Utterance segmentation

The audio files are usually very long, more than one and a half hours for a movie and about 40 to 50 minutes for an episode of TV play. The software used for operating them in this project is Praat, with which long sound files can be opened by long sound file. When the sound file is open, utterances can be selected. An utterance is defined here as a complete unit of speech, ranging from a single word to a long uninterrupted chunk of speech. In this project, one of the aims is to collect functional into nations, so the expressive completeness of an utterance is taken into consideration when segmenting it. Therefore, in most cases, an utterance corresponds to a sentence. In a movie

or TV play, there is sometimes music or noise in the background while the character is speaking. These utterances should be rejected. Only the comparatively clear sound can be picked. Fig. 1 shows two examples of utterances, with that in Fig. 1 clear and that in Fig. 2noisy. Besides waveform, judgment can be also made by observing the spectrogram. A usable utterance should meet the following two criteria. Firstly, in the wide-band spectrogram, syllable boundaries should be detectable. Secondly, in the narrow-band spectrogram, fundamental frequency and harmonics should be detectable. If an utterance meets these criteria, the syllable boundaries can be marked and its pitch can be obtained.

Fig 1: Clear sound

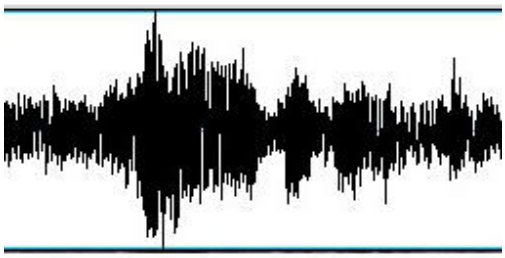

Fig 2: Noisy sound.

If an utterance is usable, it will be selected and saved as a separate wave file.

III. Assamese language:

The Assamese (অসমীয়া)[17]language basically is an Eastern Indo-Aryan language. It is the mostly used and official language of Assam. Some part of Arunachal Pradesh, Nagaland and some other northeast Indian states speaks Assamese language. A Small part of Assamese speakers can be found in Bangladesh. The sister language of Assamese languages include Bengali, Oriya, Maithili, Chittagonian, Sylhetiand Bihari languages. The Assamese language has its own script. Assamese Vowels can either be independent or dependent upon a consonant or a consonant cluster. The word Assamese is an English formation built on the same principle as Sinhalese or Canarese. The Assamese language has eight vowels, ten diphthongs, and twenty-one consonants. The vowels and consonants of Assamese language are shown below.

Vowels

	Front			Central			Back		
	IPA	ROM	Script	IPA	ROM	Script	IPA	ROM	Script
Close	i	i	ই/ি				u	u	উ/ু
Near-close							ʊ	û	ঽ
Close-mid	e	e	এ'				o	o	অ'
Open-mid	ɛ	ê	এ				ɔ	ô	অ
Open				a	a	আ			

Consonants

		Labial			Alveolar			Velar			Glottal		
		IPA	ROM	Script	IPA	ROM	Script	IPA	ROM	Script	IPA	ROM	Script
Nasal		m	m	ম	n	n	ন/ণ	ŋ	ng	ঙ/ং			
Stop	voiceless	p	p	প	t	t	ত/ট	k	k	ক			
	aspirated	pʰ	ph	ফ	tʰ	th	থ/ঠ	kʰ	kh	খ			
	voiced	b	b	ব	d	d	দ/ড	g	g	গ			
	murmured	bʱ	bh	ভ	dʱ	dh	ধ/ঢ	gʱ	gh	ঘ			
Fricative	voiceless				s	s	চ/ছ	x	x	শ/ষ/স	h	h	হ
	voiced				z	z	জ/ঝ/য						
Approximant		w	w	ৱ	l, ɹ	l, r	ল, ৰ						

IV. Information Recording

Our experiment is aimed at collecting various intonations, expressing different functions and emotions. The relevant information includes the following.

A) Type of the video

This records whether the video is a movie or a TV play.

B) Name of the movie or TV play

This records the source of the utterance.

C) Theme of the movie or TV play

D) Character

The name of the speaker is recorded.

E) Gender of the character

The gender of the character is recorded.

F) Profession of the character

This records the character's profession.

G) Relationship between the speaker and the listener

The relationship of the speaker and the listener is recorded, such as teacher to student, father to son.

IV. Pitch Extraction

In the experiment pitch extraction is also has done. As the signal to noise ratio (SNR) is not very well for some utterances, the extraction is sometimes not very well either. In this case, manual correction is done by viewing the narrow-band spectrogram.

V. Labeling

A) Syllable and prosody labeling

The speech transcription is crucial for determining voice segments within a speech database because it describes the various linguistic phenomena in the audio waveform by written text or symbols. Normally, corpus speech is labeled in different tiers. In this project, two tiers of labeling are done, that is, syllable tier and prosody tier.

B) Function labeling

The functional types are labeled in this experiment. Generally speaking, there are functionally fourmaj or types of sentences, statement, question, command and exclamation.

C) Emotion labeling

In case of movie and TV play conversations, some of the utterances are apparently not emotional, which is labeled "neutral". But in certain situations, the characters' utterances are obviously emotional. For these utterances, the emotions they conveyed will be labeled accordingly.

VI. Research on Functional Intonation

In Assamese, different functions may be expressed by different intonations. For example, the following sentence,(3) Rahit Das. It's Saturday. If the pitch in (3) falls to a very low, then it is as statement. But if the end of the pitch is not very low, the sentence is a question, or a doubt.

VII. Research on Emotional Intonation

Much of the research work has been done to examine how vocal emotions are encoded, from which we know that emotional meanings in the voice are conveyed by changes in several acoustic parameters of speech, including to fundamental frequency, duration, rhythm, and different aspects of voice quality. The fact that most of the researchers have measured changes in pitch, intensity, and speech rate implies that these parameters are critical features of emotional expressions. In particular case, a speaker's pitch level, pitch range, and speech rate appear to differentiate among discrete emotion categories in both acoustic and perceptual terms. For example, expressions of sadness tend to be produced with a relatively low pitch and slow speaking rate, whereas expressions of anger and happiness tend to be produced with a moderate or high mean pitch and fast speaking rate.

VIII. Functional and Emotional Recognition

The most usage of any speech corpora is for training a speech recognizer [1]. When we listen to any speech, we can recognize the speaker's interactive function and emotional state. Since the human speech conveys the speaker's interactive intention and emotional state. A given emotion is considered as a result of the interaction among acoustical, psychological, and physiological features. Emotion is experienced at a time when something unexpected happens, arising suddenly in response to a particular event [16]. For natural human-machine interaction, there is a requirement of machine based functional and emotional intelligence. For satisfactory responses to human interactive functions and emotions, computer systems need accurate function and emotion recognition and that correct recognition of human interactive function and emotion improves efficiency of human-machine interaction.

IX. Conclusion

Throughout this paper, the design and development of a database of functional and emotional intonation in Assamese is described. The main purpose of our database is to work as a resource for studying functional and emotional intonation and its recognition and synthesis. It is based on conversations from movies and TV plays of about 100 hours, with utterances segmented, information recorded and besides this, syllable and prosody annotation is done and pitch is extracted and manually corrected. The database contains large number of utterances and their transcriptions, pitch values, etc. by various speakers. It will be

applicable for functional andemotional intonation studying.

X. References

[1] Chotimongkol, A., K. Saykhum, P. Chootrakool, N.Thatphithakkul and C. Wutiwiwatchai, "LOTUS-BN: A ThaiBroadcast News Corpus and Its Research Applications," *Proc. of Oriental-COCOSDA 2009*, Urumqi, China, pp. 44-50.

[2] Yu, H., L., Gao, L. Guo, and J. Kou, "An Overview of Tibetan Corpus Construction," *Proc. of Oriental-COCOSDA 2010*, Nepal.

[3] Li, A., B. Xu, et. al., "A spontaneous Conversation Corpus CADCC," *Proc. of Oriental COCOCSDA 2001*, Korea.[4] Li, A., Z. Yin, et. al., "RASC863-A Chinese Speech Corpus with Four Regional Accents," *Proc. of ICSLT-o-COCOSDA*, New Delhi, India, 2004.

[5] Wang H. and J. Pan, "An Undergraduate Mandarin Speech Database for Speaker Recognition Research," *Proc. of Oriental-COCOSDA 2009*, Urumqi, China, pp. 94-99.

[6] Hu, X., R. Isotani and S. Nakamura, "Construction of Chinese Conversational Corpora for Spontaneous Speech Recognition and Comparative Study on the Trilingual Parallel Corpora, "*Proc. of Oriental-COCOSDA 2009*, Urumqi, China, pp. 56-59.

[7] Takahiro, M., T, Fukuda, H. Kikuchi and K. Shirai, "Method for Collection of Diverse Speech for Emotion Research Database," *Proc. of Oriental-COCOSDA 2010*, Nepal.

[8] Steidl, S. *Automatic Classification of Emotion-Related User States in Spontaneous Children's Speech*, Logos Verlag, Berlin, 2009.

[9] Boersma, P., "Praat, a system for doing phonetics bycomputer," *Glot International*, 5:9/10, pp. 341-345, 2001.

[10] McEnery, T. and A. Wilson, *Corpus Linguistics*, Edinburgh:Edinburgh University Press, 1996.

[11] Bird, S., and M. Liberman, "A Formal Framework for Linguistic Annotation," *Speech Communication*, 33: pp. 23-60, 2001.

[12] Li, A., "Chinese Prosody and Prosodic Labeling of Spontaneous Speech," *Proc of Speech Prosody*, Aixen-Provence, France, pp. 11-13, 2002.

[13] Cruttenden, A., *Intonation*, Cambridge: Cambridge University Press, 1997.

[14] Banse, R. and K. R Scherer, "Acoustic profiles in vocale motion expression," Journal of Personality and Social Psychology, 70(3): pp. 614-636, 1996.

[15] Mozziconacci, S., "Modeling emotion and attitude in speechby means of perceptually based parameter values," *UserModeling and User-Adapted Interaction II*, pp. 297-326, 2001.

[16] Garg, J., I. Khan, S. K. Gupta, and S. S. Agrawal,"Development of speech data base for various emotions and their recognition using Neural Network Classifier," *Proc. Of Oriental-COCOSDA 2010*, Nepal

[15] Mozziconacci, S., "Modeling emotion and attitude in speechby means of perceptually based parameter values," *User Modeling and User-Adapted Interaction II*, pp. 297-326, 2001.

[16] Garg, J., I. Khan, S. K. Gupta, and S. S. Agrawal, "Development of speech data base for various emotions and their recognition using Neural Network Classifier," *Proc. Of Oriental-COCOSDA 2010*, Nepal.

[17] Mozziconacci, S., "Modeling emotion and attitude in speechby means of perceptually based parameter values," *UserModeling and User-Adapted Interaction II*, pp. 297-326, 2001.

[18] Garg, J., I. Khan, S. K. Gupta, and S. S. Agrawal,"Development of speech data base for various emotions andtheir recognition using Neural Network Classifier," *Proc. OfOriental-COCOSDA 2010*, Nepal.

[19] www. google.com as web reference.