Emotional Speech Synthesis for Telugu

D.NAGARAJU

Reasearch Scholar, Bharatiyar University,
Coimbatoor ,Tamilanadu,India,
e-mail:dubisettynagaraju@gmail.com,

Dr.R.J.RAMASREE

Reader & Head of ComputerScience,
RSVP,Tirupati,AP, India
e-mail: rjramasree@yahoo.com

LIZZY VINODINI TALARI

M.Tech(C.S), Audisankara college of Engineering &Technology,
Gudur,A.P,India.
e-mail: lizzytalari@gmail.com

Abstract

This paper presents thorough study of emotional speech in Telugu. Emotional speech can improve the naturalness of speech synthesis system. Modern synthesizer has achieved a high degree of intelligibility, but cannot be regarded as natural sound. In order to decrease the monotony of the speech synthesis, implementation of emotional effects is known being progressively considered. Most of the recent work in this area has been focused on the Neutral voice. The continuous increase in synthetic speech intelligibility has focused the attention of the research in the area of naturalness. Emotional voice (sometimes under stress conditions) is analyzed in many papers in the last few years [1][2}

In this paper Section 1 describes Introduction , section 2 presents Phonetic Nature of Telugu language, section 3 presents emotions classifications, section 4 describes Prosodic parameters ,section 5 describes database design section 6 explains Experiments, section 7 gives Conclusion and section 8 represents References.

Keywords: Emotional Speech in Telugu; Intelligibility; Neutral Voice; Emotional Voice.

1. Introduction

1.1 *Rule based synthesis*

Rule based synthesis, also known as formant synthesis, creates speech through rules of acoustic correlates of speech sounds. Although there suiting speech sounds quite unnatural and metallic it has the advantage that many voice parameters can be varied freely. When modeling emotions this becomes every interesting. Burkhardt in his PhD thesis [3] employed a format synthesizer but took quite a different approach for generating the voice parameters. It produces artificial speech that does not sound like real human speech. Formant synthesizers usually are smaller programs than concatenate systems because they do not have a database of speech samples to refer when generating the speech.

1.2 *Concatenation synthesis*

Concatenative synthesis is based on the concatenation of segments of recorded speech. Generally, this form of synthesis produces the most natural sound. However, variation in human speech and the techniques used for segmenting the waveforms sometimes reduce the naturalness of the output. (Komshian and Bunnell, 1998). The diaphone concatenation synthesizer focuses from the middle of one sound to the middle of the next sound. The number of diphones depends on the language. Different languages have different number of diphones. Diphone concatenation produces synthesized speech by combining together related diphone stored in database that matches the input string. The diphone concatenative synthesizer refers to an existing database. Some current diphone concatenative synthesizers are PSOLA, MBROLA and Festival.

2. Phonetic nature of Telugu Language

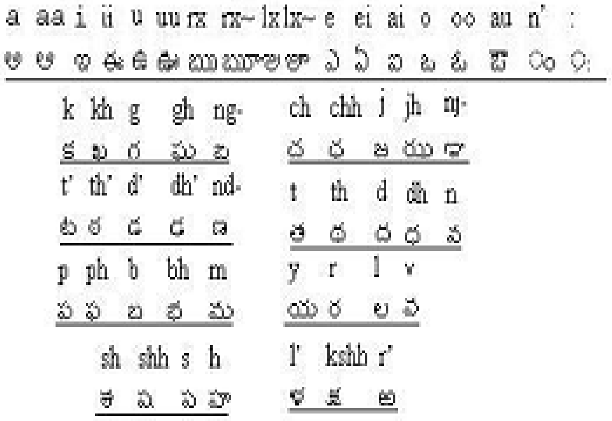
Building a new language voice[4] requires addressing the entire issues specific to that new language. The natural language processing module in a TTS system deals with the production of a correct phonetic and prosodic transcription of input text. It involves analysis of the text and generation of prosodic parameters. The output of this module is given to the waveform generation module, which transforms this symbolic information into speech.

2.1 *Nature of Telugu Language Script*

The scripts in Indian languages have originated from the ancient Brahmi script. The basic units of the writing system are referred to as Aksharas. The properties of Aksharas are as follows: (1) An Akshara is an orthographic representation of a speech sound in an Indian language; (2) Aksharas are syllabic in nature; (3) The typical forms of Akshara are V, CV, CCV and CCCV, thus have a generalized form of C*V. The shape of an Akshara depends on its composition of consonants and the vowel, and sequence of the consonants. In defining the shape of an Akshara, one of the consonant symbols acts as pivotal symbol (referred to as semi-full form). Depending on the context, an Akshara can have a complex shape with other consonant and vowel symbols being placed on top, below, before, after or sometimes surrounding the pivotal symbol (referred to as half-form). Thus to render an Akshara, a set of semi-full or half-forms have to be rendered, which in turn are rendered using a set of basic shapes referred to as glyphs. Often a semi-full form or half-form is rendered using two or more glyphs, thus there is no one-to-one correspondence between glyphs of a font and semi-full or half-forms.

2.2 *Telugu Language Script*

To handle diversified storage formats of scripts of Indian languages such as ASCII based fonts, ISCII (Indian Standard code for Information Interchange) and Unicode etc, it is useful and becomes necessary to use a meta-storage format. A transliteration scheme maps the Aksharas of Indian languages onto English alphabets and it could serve as meta-storage format for text-data. Since Aksharas in Indian languages are orthographic representation of speech sound, and they have a common phonetic base, it is suggested to have a phonetic transliteration scheme such as IT3. Thus when the font-data is converted into IT3, it essentially turns the whole effort into font-to-Akshara conversion. The following base-map table provides the mapping basic between the glyphs of the font-type to the Aksharas represented in IT3 transliteration scheme.



a	aa	i	ii	u	uu	rx	rx~	lx	lx~	e	ei	ai	o	oo	au	n'	:
అ	ఆ	ఇ	ఈ	ఉ	ఊ	ఋ	ౠ	ఌ	ౡ	ఎ	ఏ	ఐ	ఒ	ఓ	ఔ	ం	ః

k	kh	g	gh	ng-	ch	chh	j	jh	nj-
క	ఖ	గ	ఘ	ఙ	చ	ఛ	జ	ఝ	ఞ
t'	th'	d'	dh'	nd-	t	th	d	dh	n
ట	ఠ	డ	ఢ	ణ	త	థ	ద	ధ	న
p	ph	b	bh	m	y	r	l	v	
ప	ఫ	బ	భ	మ	య	ర	ల	వ	
sh	shh	s	h		l'	kshh	r'		
శ	ష	స	హ		ళ	క్ష	ఱ		

Fig. 1. Telugu Script

3. Emotions classification

3.1 *Classifying Emotion*

Human emotion falls into various categories and there are many ways to classify human emotion. This dissertation introduces two approaches which are categorical and dimensional approach.

3.2 *Categorical Approach*

The Categorical approach attempts to define specific categories or types of emotions. This approach suggests that there are a number of basic emotions (between three to more than twenty). Emotions that are normally classified as basic emotions are disgust, anger, happiness, sadness and fear (Cahn, 1989; Davitz, 1964; Fairbanks and Pronovost, 1939). Some of basic emotions are discussed below.

3.1.1**.** *Happiness*

Happiness or joy can be described as a feeling of gladness over a certain event or circumstances and it makes a person feel good and excited. We get this feeling when something really touches our hearts and we are actually very grateful that such a thing has happened. Most people portray happiness with smile and laughter. However there are other reactions to portray a happy emotion. Some people may even reach a stage of happiness that they actually weep while smiling and there are people who react to happiness by cheering and hugging.

3.1.2**.** *Anger*

Anger is an emotional state of aggression, and its expression conveys messages about hostility, opposition, and potential attack. Anger is a common response to anger expressions, thus creating a positive feedback loop and increasing the likelihood of dangerous conflict. Although frequently associated with violence and destruction, anger is probably the most socially constructive emotion as it often underlies the efforts of individuals to shape societies into better, more just environments, and to resist the imposition of injustice and tyranny.

3.1.3**.** *Sadness*

Sad expressions are often conceived as opposite to happy ones, but this view is too simple, although the action of the mouth corners is opposite. Sad expressions convey messages related to loss, bereavement, discomfort, pain and helplessness.

3.1.3**.** *Fear*

Fear is human response to an external threat and usually produce disturbance shown by a person’s action and body changes. Feelings of fear are varied in both type and intensity. Fear may be trigged by internal or external events, conditions, or situations that signal danger. The threat may be physical or psychological.

3.3 *Dimensional Approach*

The dimensional approach conceptualizes emotion as having two or three basic underlying dimensions along which the entire range of human emotions can be arranged (Guerrero et al., 1998). The most common dimensions are valence (which ranges from happiness to sadness) and arousal (which ranges from calm to excitement). The third less often mentioned dimension is dominance (ranging from in control to out of control).

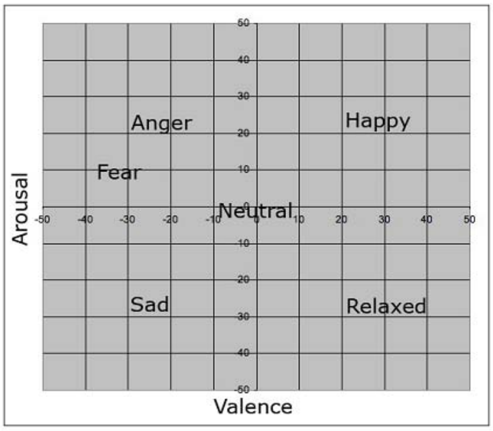


Fig. 1. Types of Emotions

4. Prosodic Parameters

The following parameters can vary to get different emotions and Naturalness of Speech.

Table 1. Prosodic Parameters

Pitch	perceived frequency of sound
Pitch range	specifies the range over which these variations may occur
Fundamental frequency F0	the lowest frequency of a periodic waveform
Intonation	variation of pitch
Pitch contour	a function or curve that tracks the perceived pitch of the sound over time
Jitter	deviation in or displacement of some aspect of the pulses in a high-frequency digital signal.
Vocal effort	is a quantity varied by speakers when adjusting to an increase or decrease in the communication distance.

5. Emotional Database

For the analysis purpose, our emotional speech database consists of 18 sentences and 10 words spoken by a male speaker and female speaker in three speaking styles anger, happiness and sadness. The words and sentences are including all the frequently occurring combinations of vowels and consonants. The selected phonemes selected so a represent the different categories as given in the linguistic structure of the alphabet set. Signals were recorded in an acoustically treated room, a high quality directional micro phone was used, and wave forms were digitally acquired for the following analysis, these signals were down sampled to 16KHz.

6. Experiments

6.1 *Perception Test of Recorded Human Speech*

The main objective of the perception test is to validate the recorded voice for recognition of emotive elements. The perception test that was conducted for this research involved twenty five people from various backgrounds. This perception test was divided into tests. In test one, listeners listened to a list of randomized recorded human voice and described the emotion without any list of choices. Listeners were also required to rate the effort taken to recognize the emotional elements in each sentences from the scale of one to five (1 for more effort; 4 for least effort). For the second test, listeners were required to choose the emotions of the recorded voice from a list of three emotions: happiness, sadness and anger.

The result from test reveals that emotions of high arousal were better recognized then emotions with low arousal (Fairbanks and Hoaglin, 1941). The result from the first test was summarized in table below.

Table 2. Recognition of Emotions

Sentences	Happy(%)	Angry(%)	Sad (%)	Other(%)	Effort rating
Happy	75.34	2.26	1.00	21.4	3.144
Angry	2.54	79.00	2.33	16.13	3.355
Fear	3.33	5.00	70.00	21.67	3.133

The overall result of test one shows that the average recognition rate for happy sentences was 75.34% and the average effort taken was 3.144. Happy was commonly confused for boasting, excited and proud. For angry sentences, the average recognition rate was 79.00% and the average effort taken to recognize this emotion was 3.355. Fear sentences have the lowest recognition rate among the four emotions at 70.00%. The effort rate for fear sentences was 3.133. The confusion rate of sadness and fear was much greater than anger and happiness.

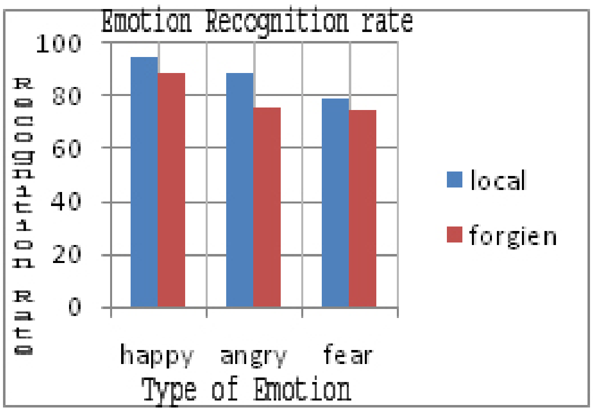


Chart 1: The Average Recognition Rate for Both Local & Foreign Listeners

The perception test also shows that recognition rate for the emotion increases when listeners are given a list of choices. The average effort rate for happiness is 3.57, anger is 3.65, and fear is at 2.93. Listeners placed less effort to recognize the happiness and anger but more effort is needed to recognize fear. Figure shows the result of emotion recognition rate for both local and foreign listeners. Figure shows the result of emotion effort rate for both local and foreign listeners.

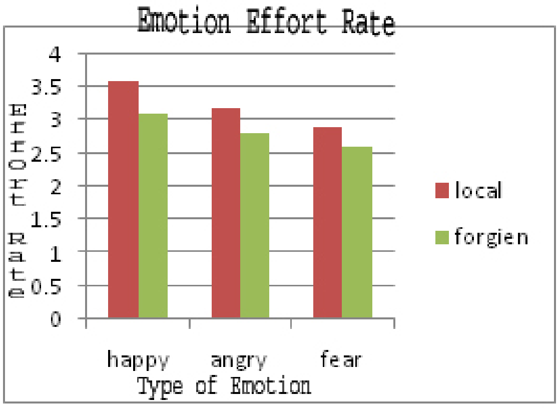


Chart 2: The Average Effort Rate for Both Local & Foreign Listeners

7. Conclusion

Lot of research is ongoing in the area of Telugu Emotional speech synthesis. In the previous system we have implemented a good quality speech synthesis system for Telugu. In this paper we have briefly discussed the development of Emotional speech synthesis system for Telugu. Still the research is ongoing in this area. We hope that, shortly we will release a good quality emotional speech synthesis system for Telugu for both Commercial and academic purpose.

8. References

[1] Scherer K.R. (1996) "Adding the affective dimension: a new look in speech analysis and synthesis" in Proceedings of ICSLP'96.

[2] Amir N. and Ron S. (1998) "Towards anautomatic classification of emotions in speech" in Proceedings of ICSLP'98.

[3] Burkhardt,F.(1999).Simulation emotional erSprechwe is emit Sprach synthe sever fahren. PhDthesis,TUBerlin

[4] FestVox: A. Black and K. Lenzo, "Building voices in the Festival speech synthesis system", www: http://festvox.org/festvox/festvox toc.html

[5] FestVox: A. Black and K. Lenzo, "Building voices in the Festival speech synthesis system", www: http://festvox.org/ festvox/ festvox toc.html

[6] F. Burkhardt, Simulation emotionaler Sprechweise mit Sprachsyntheseverfahren Dissertation an der TU-Berlin, Shaker Verlag 2001

[7] C.M. Lee and S Narayan, "Towards Detecting Emotion in Spoken dialogs," IEEE Trans. Speech and Audio Processing, Vol.13 no.2,pp. 2993-303,2005.

[8] Pascal van Lieshout, Ph.D. "PRAAT", Oral Dynamics Lab V. 4.2.1, October 7, 2003.

[9] AlanW. Black and Paul Taylor, "Automatically clustering similar units for unit selection in speech synthesis.," in Proceedings of Eurospeech, vol. 2, pp. 601–604, 1997.

[10] E. Veera Raghavendra, B. Yegnanarayana, and Kishore Prahallad, "Speech synthesis using approximate matching of syllables," in Proceedings of IEEE SLT Workshop, 2008, pp. 37–40.