Synthesis of Speech with Emotion

R. Murali shankar and A. G. Ramakrishnan
Department of Electrical Engineering
Indian Institute of Science
Bangalore 560 012, INDIA
e-mail: {sripad,ramkiag}@ee.iisc.ernet.in

Abstract

This paper describes the methodology proposed by us for synthesizing speech with emotion. Our work starts with the pitch synchronous analysis of single phoneme utterances with natural emotion to obtain the linear prediction (LP) parameters. For synthesizing speech with emotion, we modify the pitch contour of a normal utterance of a single phoneme. We subsequently filter this signal using the LP parameters. The proposed technique can be used to improve the naturalness of voice in a text-to-speech system.
Keywords: speech synthesis, emotion, pitch contour, DCT, LP filtering.

1 Introduction

Synthesized speech is mainly distinguished by a lower intelligibility, a not natural prosody and lack of expressiveness. These are the important drawbacks for computer generated speech. There are two different ways to synthesize emotional speech. The first methodology is based on a text-to-speech converter that can generate speech from carrier sentences. The emotional expression in this speech is achieved modifying its prosody appropriately. The other method uses recorded speech with neutral prosody to turn it into emotional speech with corresponding prosody.

In the study of emotion in speech, it is supposed that, voice suffers acoustical changes caused directly due to the physiological alterations of the human body when a person has a strong feeling [1,2,3]. These changes also depend on the language used. In spite of this supposition, we think that it is convenient to do a study without differentiating between linguistic and non-linguistic processes. This is done by considering emotional speech as a united system that comprises both the cultural influence of the language and physiological mechanism of emotion. On the other hand, we have considered that emotions suffer a dynamic evolution in time with a variable duration [4]. Also the acoustical features that determine each emotional state are independent of the word uttered, or the language. Then, to do the study, we have analyzed suprasegmental voiced patterns, because they jointly hold the features of the language and the acoustical features dependent on each emotional state. The five primary emotions are Anger, Joy, Fear, Sadness and Disgust followed by secondary emotions. The term "emotion" is used to describe any subjectively perceivable and identifiable physiological and psychological state.

The most commonly referenced vocal parameters in the emotion literature are pitch (i.e., fundamental frequency, both its average value and range), duration, intensity and the undefined term "voice quality". Pitch, being correlated closely to the activity level, is often used to differentiate between emotions. Vocal indicators of emotional expression are, principally, the pitch, intensity, and voice quality. While pitch is important in emotional expression, voice quality is more important in differentiating between sudden variations in emotion [3]. Variations in these same parameters also convey stress information [5]. As stress (at both the sentence and word level) is independent of the speaker's emotion, the changes caused by emotion must be smaller in amplitude and superimposed on those values of underlying neutral utterances. Cowan [6] stated that unemotional speech has a narrow pitch range compared to emotional speech, with pitch tending to be normally distributed about the average pitch level. He also commented that every emotion might be expressed in varying degrees of force. Emotion can be explained by noting that fewer and shorter interruptions occur in speech flow as a speaker becomes more excited, long pauses being eliminated altogether. Moreover the "speech rate" is not related to the degree of emotionality since a relation is noticed between the percentage of phonated time and degree of emotional expression. Emotional utterances are divided into "active" and "passive" groups. Davitz [7] characterized pas-

sive emotions as those with low speech rate, low volume, low pitch, and more resonant timbre (voice quality). Active emotions are characterized by their high speech rate, high loudness, high pitch, and blazing timbre. The gentler emotions are closer to music and have song-like qualities. Studies of the effects of emotion on the acoustic characteristics of speech have shown that average values and ranges of fundamental frequency (F0[s1]) differ from one emotion to another [8]. A preliminary study [9] showed that the most sensitive indicator of emotion was the contour of F0 throughout an utterance. Since only certain aspects of the F0 contour carry information regarding the linguistic content of a message, considerable latitude is possible in the variation of F0. Subject to certain constraints, a speaker is free to use changes in F0 to convey nonlinguistic information, such as his emotions. Furthermore, the fundamental frequency can undergo variations that may not be intended or be under overt control of the speaker, and hence may provide an indication of the speaker's emotional state.

Other relevant physiological effects of certain emotions are dryness of the mouth, and tremor and disorganization of motor response. These effects have an influence on various components of the speech system, including the larynx, which is directly involved in the control of F0. Muscle activity in the larynx and the conditions of the vocal cords are likely to have a direct influence on sound output and in particular, on the fundamental frequency. The reason is that the vibrating vocal cords have a direct effect on the volume velocity through the glottis, whereas the other muscles and vocal-tract components simply shape the resonant cavities for the sound generated at the vocal cords. Thus, any analysis of the speech signal that reflects vocal-cord activity must characterize the physiological changes brought about by the emotional state of the speaker. Physiological changes such as increased subglottal pressure, excessive dryness or salivation, and decreased smoothness of motor control can have an influence on the waveform of the pulses from the vocal cords, as well as on their frequency. For example, increased subglottal pressure generally gives rise to a narrowing of individual glottal pulses, and hence to a change in spectrum of the pulses. Thus, we conclude that the pitch-contour and spectral variations are sufficient enough to synthesize emotion.

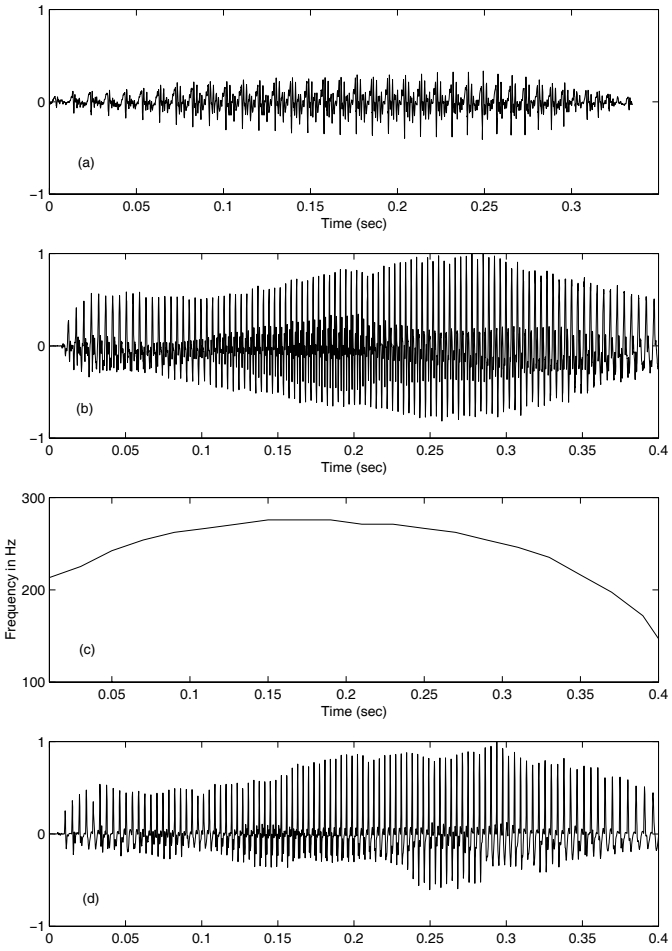

Figure 1: (a) Emotionless utterance, (b) Utterance with emotion (c)Pitch contour of (b), and (d) Synthesized Signal

2 Method

In the first step, we carry out pitch marking on an emotional speech segment, instead of utilizing the general pitch measurement algorithm, which approximates the pitch contour. Pitch marking enables us to follow the pitch dynamically and is very important because accurate pitch marking results in better synthesized output.

The parameters that we have considered are pitch contour and spectral information, within the pitch interval. We are using this information to modify a segment of emotionless signal and then repeat the segment with varying lengths so as to get the pitch contour of the emotional sequence. The spectral information of emotional signal is obtained by pitch synchronous LP analysis. Transform domain as well as time domain filtering incorporates this information in the emotionless segment of varying length. The length of the segment is chosen on the basis of maximum pitch period of the emotional signal, which in-turn corresponds to the maximum IDCT (inverse discrete cosine transform) length. Therefore, the IDCT length is always lesser than or equal to the DCT length. Here, we drop

the higher DCT coefficients (utilizing the energy compaction property of DCT) depending on the difference in length between DCT and IDCT. Thereby we are able to adjust the pitch period dynamically depending upon the pitch contour of the emotional signal. This, however, removes some high frequency information. The quality of vowels are not affected by dropping the higher DCT coefficients because a major set of vowels are dependent upon formant frequencies F1and F2, which lies in lower frequency region of the spectrum. Energy normalization is achieved by matching the energy of each filtered segment to that of the corresponding segment of the emotional speech. Similarly, we process it for the entire pitch contour of emotional speech. Finally we concatenate all of the processed segments to complete the synthesis of the speech, full of emotion.

3 Results and Discussion

We have analyzed utterances expressing "pain" and "tiredness", and used the above algorithm to synthesize both of these emotions. The original "pain" signal analyzed is shown in Fig. 1b and its pitch contour, in Fig. 1c. Fig. 1a shows an emotion-free utterance, which is used as the input for synthesis. The synthesized signal is shown in Fig 1d. Fig. 3 & 4 compare the first three formants of the synthesized signal with those of the original signal. We find that F1 and F2 are similar and the formant F3 is different from the original signal.

Fig. 2a shows the original "tiredness" signal, and Fig. 2b, the synthesized one. Fig. 2c shows the pitch contour of the signal, and Fig. 2d gives the formant tracks for the synthesized signal. Perceptual evaluation by several volunteers has shown that the synthesized signals for both the cases are acceptably good, and do evoke the intended emotions in the listener's minds.

4 Conclusions

This work can be extended to synthesize words with emotions, which has potential application in a Text-to-Speech synthesis system to generate natural sounding speech. The pitch contour could be stored by approximating it with polynomial fitting techniques and storing only the coefficients of the polynomial.

5 References

1. K. R. Scherer,"Methods of research on vocal communication: paradigms and parameters", in K.R. Scherer and P. Ekman Eds., Handbook of Methods in non verbal behavior research, Cambridge Univ. Press, 1982.

2. K. R. Scherer, "Vocal affect signaling: A comparative approach", in J.S. Rosenblatt, C. Beer, M.C. Busnel and P.J.B. Slater Eds., Advances in the Study of Behavior (Vol. 15). New. York, Academic Press, 1985.

3. K. R. Scherer, "Vocal affect expression: a review and a model for future research", Psichological Bulletin, Vol. 99, pp. 143-165, 1986.

4. J. M. Reeve, Motivacin y emocin, McGraw-Hill /Interamericana de Espaa, S.A, Aravaca (Madrid), 1994.

5. R. Ortleb, "An objective study of emphasis in oral reading of emotional and unemotional material", Speech monograph, Vol. 3, 56-58, 1937.

6. M. Cowan, "Pitch and intensity characteristics of stage speech", Arch. Speech, Suppl. to Dec. issue, 1936.

7. J. R. Davitz, "Personality, Perceptual and Cognitive Correlates of Emotional Sensitivity", The Communication of Emotional Meaning. New York, McGraw-Hill, pp.101-112, 1964.

8. G. Fairbanks and W. Pronovost, "An experimental study of pitch characteristics of voice during the expression of emotion", Speech Monograph, 6, 87-194, 1939.

9. C. E. Williams, K. N. Stevens and M. L. Hecker, "Acoustical manifestations of emotional speech", J. Acoust. Soc. Amer., Vol. 47, 66(A), 1970.

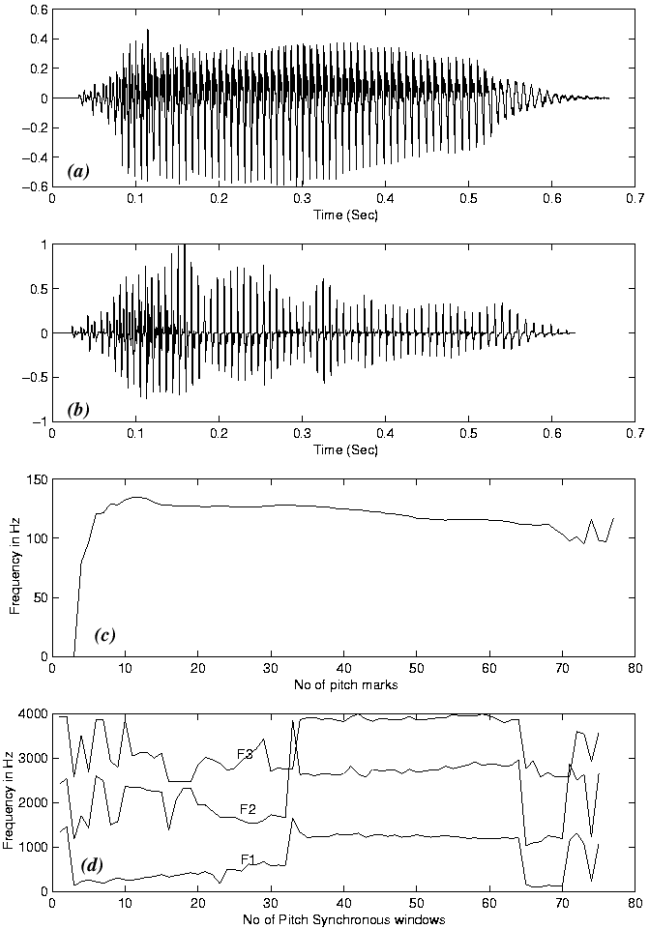

Figure 2: (a) Emotional ("Tiredness") Speech, (b) Synthesized Speech, (c) Pitch Contour of the emotional signal, and (d) Formants of the synthesized speech.

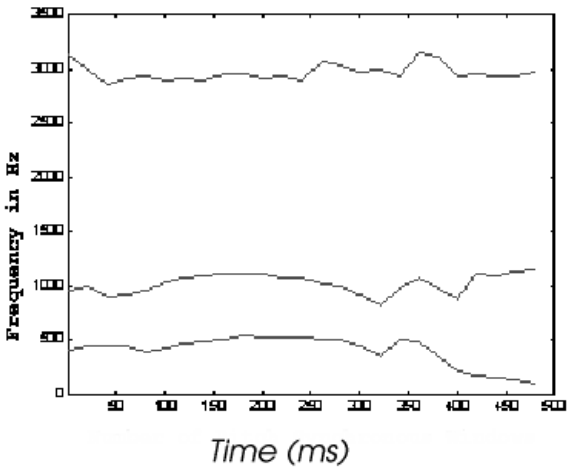

Figure 3: Formant tracks for emotional signal (pain).

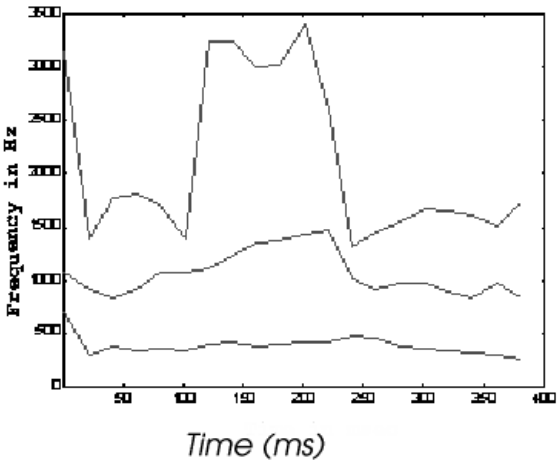

Figure 4: Formant tracks for synthesized signal (pain).