Relative Importance of Different Components of Speech Contributing to Perception of Emotion

P. Gangamohan[1] *V. K. Mittal*[2] *and B. Yegnanarayana*[3]

International Institute of Information Technology, Hyderabad

[1]gangamohan.p@students.iiit.ac.in [2]vinay.mittal@iiit.ac.in, [3]yegna@iiit.ac.in

Abstract

The objective of this study is to understand the relative importance of different components of speech that contribute to perception of emotion in speech. The four components considered in this study relate to the vocal tract system, excitation source and suprasegmental (pitch and duration) information. For this study, data collected from an artist, producing speech with different emotions is used. A flexible analysis-synthesis tool is used to modify the parameters of speech in a desired manner. Results of subjective studies show that all the four components are important in perceiving emotion in an utterance in comparison to the corresponding neutral utterance. Individually, the pitch contour seems to be the dominant component, and the duration seems to play less a significant role. It is also interesting to note that the importance of these components vary in perception for different types of emotions.

Index Terms: speech prosody, speech analysis, speech synthesis, emotion conversion, dynamic time warping.

1. Introduction

The objective of this study is to determine the components of speech that contribute to the perception of emotion in an utterance. This also helps to identify out the components of speech that are emphasized by a person in producing speech under a given emotional state. Some earlier studies in similar direction are [1, 2, 3]. For this study, a controlled set of experiments are conducted to modify the speech of the same sentence uttered in neutral mode and in one of the emotional modes. The speech data collected by an artist producing an utterance in different emotional states is used [4]. The utterance in the neutral state is modified synthetically to an utterance of an emotional state by varying different components (features) of speech in the utterance. This is accomplished by using a flexible analysis-synthesis tool (FAST) developed recently [5]. The tool enables conversion of an utterance in one emotional state to an utterance of neutral state. This tool is also used to convert a neutral utterance to an emotional utterance.

The main feature of the FAST is to match two utterances of the same sentence by the same person to determine the warping path [6]. With the help of the warping path, it is possible to know how the different components of speech get modified. In this study four components of speech are considered for modification. They correspond to speech production system characteristics and suprasegmental information. The results of modification of these components are evaluated using subjective studies. Preliminary results of these studies are reported in [5]. This paper presents more detailed analysis of the results to show the significance of different components of speech in perceiving emotion characteristics in an utterance.

The paper is organized as follows: The four components of speech used for modification are discussed briefly in Section 2. In Section 3 the flexible analysis-synthesis tool is described briefly. In Section 4 the experiments conducted to convert a speech utterance from neutral state to an emotional state, and vice versa, using FAST are described. The results of subjective studies on the effects of different components on perception of emotion are discussed in Section 5. Section 6 gives a summary and scope for further studies.

2. Components of speech for study of emotion characteristics

Speech is produced by the excitation of the time-varying vocal tract system with time-varying excitation. The speech signal carries information about the dynamic vocal tract system, the excitation source, the duration of different sound units and intonation. While the vocal tract system and excitation characteristics reflect mainly the inherent characteristics of the speech production system, the suprasegmental information consisting of duration and intonation is mostly due to acquired characteristics of an individual over a period of time. Thus the duration and intonation (i.e., prosody features) relate mainly to the behavioural traits of the speaker. The duration information reflects the speaking style and speaking rate, and the intonation component should reflect mostly the emotion.

The time-varying vocal tract information is represented by the linear prediction coefficients (LPCs) obtained by using linear prediction (LP) analysis on each frame of 20 msec for every 10 msec. The excitation information is represented through LP residual signal within each glottal cycle. The LPCs can be changed as dictated by the warping path to modify the characteristics of the vocal tract system. Likewise the LP residual within each glottal cycle can be replaced in a desired fashion, to suit the requirements of the target characteristics.

The excitation characteristics such as pitch contour and duration information are derived from the epoch locations and strength of impulses using the zero frequency filtering (ZFF) method [7]. The method involves passing the speech signal through a cascade of two ideal digital resonators located at 0 Hz. By removing the trend in the output, the characteristics of the impulse-like excitation can be derived. The trend removal operation involves subtracting the local mean computed at every sampling instant. The local mean is computed over an interval corresponding to about 1.5 times the average pitch period. The mean subtracted signal is called zero frequency filtered (ZFF) signal. The negative to positive zero crossings of the ZFF signal correspond to the instants of significant excitation of the vocal tract. These instants, called epochs, occur at the instants of glottal closure in each glottal cycle of voiced speech. The slope of

the ZFF signal at an epoch gives an indication of the strength of the impulse-like excitation [7, 8]. The strength of epochs or the energy of the ZFF signal can be used to determine the regions of voiced and unvoiced segments. The interval between successive epochs gives the value of the instantaneous pitch period (T_0) or the instantaneous fundamental frequency ($F_0 = 1/T_0$).

Duration features of the source component of an utterance are modified frame-wise as dictated by the warping path, instead of modifying the relative durations of sound units such as syllables, words etc. This is because no a priori knowledge of relative significance of the sound units is available.

3. Flexible Analysis-Synthesis Tool (FAST)

We will briefly describe this tool, as it is used in this study to generate synthesized speech by modifying the characteristics of the source utterance [5]. The tool is used to convert an emotion utterance into neutral one, and vice versa. The tool uses a dynamic time warping (DTW) algorithm to match two utterances. The utterances are represented by a sequence of vectors, each vector representing the characteristics of the vocal tract system for a given analysis frame. In this study we use a 20-dimensional linearly weighted LP cepstral coefficients (wLPCC) vector, derived using a 10^{th} order LP analysis on a segment of 20 msec taken for every 10 msec.

Time alignment of the source utterance with the target utterance is carried out using the DTW algorithm [6]. The algorithm consists of matching two speech utterances (X and Y) of length M and N frames, represented as two sequences of vectors $(\mathbf{x_1}, \mathbf{x_2}, \ldots, \mathbf{x_M})$ and $(\mathbf{y_1}, \mathbf{y_2}, \ldots, \mathbf{y_N})$, respectively. Each of these vectors $\mathbf{x_i}$ and $\mathbf{y_j}$ correspond to the wLPCC vectors for the i^{th} and j^{th} frames, respectively, of the source and target utterances. The DTW algorithm is based on the principle of dynamic programming, in which the optimal path to the point $(i,\ j)$ in a two dimensional matrix must pass through either the point $(i-1,\ j)$ or $(i-1,\ j-1)$ or $(i,\ j-1)$. The minimum accumulated distance to the point $(i,\ j)$ is then given by

$$\begin{aligned} D(i,j) &= d(i,j) + min\{D(i-1,j), \\ &\quad D(i-1,j-1), D(i,j-1)\}, \end{aligned} \tag{1}$$

where $d(i,\ j)$ is the distance between the vectors $\mathbf{x_i}$ and $\mathbf{y_j}$. In this study $d(i,\ j)$ is the Euclidean distance between two wLPCC vectors. The algorithm recursively computes this distance to determine the minimum accumulated distance to the point $(M,\ N)$.

The path corresponding to the minimum accumulated distance is called the *warping path*, and it gives the matching frames between source and target utterances. Two warping paths are derived for each pair of utterances. Warping path 1 (say *WP1*) is where all frames of the target utterance are used. In this case the duration also is modified automatically. Warping path 2 (say *WP2*) is where all frames of the source utterance are used. The warping path gives the pitch periods and residual signals of the target utterance corresponding to the pitch periods and the residual signals of the source utterance, respectively. Thus the pitch contour of the source utterance and the matching pitch contour of the target utterance can be obtained. Note that if the pitch contour of the source (without mapping) and the pitch contour of the corresponding target utterance are taken, then the duration of the source utterance is not modified. Only the pitch contour is modified.

Table 1: *Criterion for 'similarity score', used in perceptual listening test to judge the similarity/dissimilarity between two utterances.*

Perceptual difference between 2 utterances	Similarity score
sounds very much similar	5
sound some-what similar	4
sounds little different, little similar	3
sounds some-what different	2
sounds very much different	1

4. Modification of different components

The following experiments are conducted:-

E1: Pitch modification: To modify the pitch alone of the source utterance to that of the target utterance, the pitch of the frames in the source utterance is replaced with the pitch of the target frames according to the path WP2. Note that in this case all frames of the source utterance are used. From the new pitch contour, new epoch sequence and the LP residual sequence are obtained as in [8].

E2: Duration modification: To modify the duration alone of the source utterance to match that of the target utterance, each target frame is replaced by the corresponding source frame information using the warping path information in WP1. The new pitch contour on the target sequence frames is used to derive the new epoch sequence, and the LP residual sequence is obtained as in [8].

E3: LPCs modification: To modify the LPCs alone, the LPCs for each frame of the source sequence is replaced with the LPCs of the corresponding frames of the traget utterance as per the warping path WP2. The LP residual signal of the source utterance is used to excite the new time-varying LPCs for each frame.

E4: Pitch and duration modification: For this, the warping path WP1 is used. The epoch sequence is derived using the pitch contour of the target utterance. The LPCs are those of the source utterance.

E5: Duration and LPCs modification: This is done similar to separate duration and LPCs modifications using the warping path WP1.

E6: Pitch and LPCs modification: This is done similar to separate pitch modification using the warping path WP2.

E7: Pitch, duration and LPCs modification: This is done similar to separate duration modification using the warping path in WP1.

E8: Pitch, duration, LPCs and residual modification: This case is similar to (E7) above, except that the LP residual between two epochs is replaced by an Liljencrants-Fant (LF) model [9], where the LF model parameters are modified proportional to the interval between the epochs.

5. Results and discussion

The emotion database, named Simulated Emotion Speech Corpus, collected by the Indian Institute of Technology (IIT) Kharagpur (IIT-KGP SESC) India is used for studying the significance of different components of speech for the perception of speech. The database in Telugu (an Indian language) consists of total 12000 utterances spoken by radio artists. Each of the 10 speakers (5 male and 5 female) recorded utterances of 15 different sentences, each spoken in 8 different emotions, and repeating these utterances in 10 separate sessions of recording.

Table 2: *Average similarity scores for each emotion: Conversion of neutral (source) utterance to emotion (target) utterance. Comparison pairs are: (a) Neutral and emotion (b) Neutral and synthesized emotion, and (c) Original emotion and synthesized emotion*

Experiment–>		E1		E2		E3		E4		E5		E6		E7		E8
Comparison–>	(a)	(b)	(c)	(b)	(c)	(b)	(c)	(b)	(c)	(b)	(c)	(b)	(c)	(b)	(c)	(c)
Column # –>	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Anger	1.7	3.3	3.4	3.3	2.4	4.1	2.2	2.3	4.0	2.2	3.9	3.3	2.5	1.9	4.0	4.3
Compassion	1.8	2.6	3.0	4.1	2.3	3.8	2.5	2.8	3.5	2.5	3.9	3.4	2.3	2.6	4.0	4.0
Disgust	1.9	2.7	2.4	3.0	2.4	3.6	2.9	2.4	2.5	2.7	3.9	2.7	3.4	2.3	4.1	3.5
Fear	1.5	2.3	2.9	3.3	2.1	3.1	2.0	2.2	2.9	2.0	3.5	2.8	2.2	1.9	3.7	3.3
Happy	1.1	2.0	3.5	4.2	1.9	4.2	2.0	2.9	2.7	2.0	4.0	3.1	2.6	1.8	3.6	3.5
Sarcastic	1.3	3.2	2.5	3.4	2.1	3.4	2.1	2.2	3.1	2.1	3.4	2.1	3.4	1.7	3.9	4.0
Surprise	1.6	2.6	2.3	2.3	2.9	3.6	1.8	1.8	3.8	2.3	2.7	2.3	3.3	1.7	4.0	3.7

Table 3: *Average similarity scores for each emotion: Conversion of emotion (source) utterance to neutral (target) utterance. Comparison pairs are: (a) Neutral and emotion (b) Emotion and synthesized neutral, and (c) Original neutral and synthesized neutral*

Experiment–>		E1		E2		E3		E4		E5		E6		E7		E8
Comparison–>	(a)	(b)	(c)	(b)	(c)	(b)	(c)	(b)	(c)	(b)	(c)	(b)	(c)	(b)	(c)	(c)
Column # –>	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Anger	1.7	3.3	3.2	4.5	2.5	4.2	2.3	2.5	3.8	2.5	4.0	3.6	2.4	2.7	4.4	4.3
Compassion	1.8	2.3	2.5	4.3	1.7	3.6	2.0	2.2	2.1	2.3	2.9	3.6	2.0	2.1	3.2	3.7
Disgust	1.9	4.4	1.5	4.2	1.4	3.7	2.0	3.5	2.2	2.8	2.7	2.5	3.0	1.9	3.5	3.7
Fear	1.5	3.0	1.9	4.3	1.7	4.3	1.7	2.8	2.1	2.8	1.9	3.5	1.8	2.3	2.7	3.3
Happy	1.1	2.4	3.2	4.3	2.2	4.2	1.8	2.5	3.3	2.0	3.1	3.2	2.2	2.4	3.2	3.7
Sarcastic	1.3	3.4	2.1	3.6	1.8	3.5	2.2	2.7	3.0	2.2	3.4	2.4	3.0	1.7	3.8	3.8
Surprise	1.6	2.9	1.8	2.3	1.7	3.8	1.8	2.4	1.6	1.8	3.0	2.5	2.5	2.0	3.1	3.5

Thus in this case utterances of the same sentence from the same speaker are available in neutral (normal) mode and in 7 different emotions (namely anger, compassion, disgust, fear, sarcastic, happy and surprise).

From this database we consider neutral and 7 emotion utterances of a sentence by a male speaker for this study. Two warping paths are obtained for each pair of utterances. Each pair consists of one neutral utterance and one emotion utterance. Thus there are seven pairs of utterances used in this study. For each pair, two sets of studies are made: Set 1: Modification of neutral to emotion. Set 2: Modification of emotion to neutral. For each set all the 8 experiments (E1 to E8) listed in Section 4 are conducted. The synthesized speech obtained after modification is used for subjective evaluation. The evaluation is carried out through listening tests with 10 student listeners from the Speech and Vision Lab at IIIT Hyderabad. Each subject was given a set of three pairs of utterances to give similarity scores as per the criterion given in Table 1. A score of 5 indicates that both the utterances in a pair sound similar, whereas a score of 1 indicates that both the utterances are different. The results are given in terms of average score (over the 10 subjects) for each experiment. In the experiments for the studies of Set 1, three pairs of utterances are used for comparison: (a) original (source) neutral & original (target) emotion, (b) original (source) neutral & synthesized emotion, and (c) original (target) emotion & synthesized emotion. Ideally, for cases (a) and (b) the similarity scores should be low, and for case (c) the similarity scores should be high, if that component of speech contributes significantly for the perception of emotion.

Table 2 gives the results of subjective evaluation for the experiments for the Set 1, namely, modification of neutral (source) to emotion (target). The entries in column 1 give low scores for all emotion categories. The scores in column 2 should be low, but these are higher indicating that the synthesized emotion by pitch modification did not produce the effect of target (emotion) significantly. The scores in column 3 are slightly higher than those in column 2, except for the emotions 'disgust', 'sarcastic' and 'surprise' This indicates that pitch modification alone captures only a small amount of emotion characteristics. Columns 4 and 5 correspond to duration modification. The results in column 5 indicates that duration modification does not bring out the emotion characteristics well, when compared to original emotion utterance. This is also indicated well by the somewhat larger scores in column 4, which means that even after duration modification the synthesized speech sounds more like source (neutral) utterance.

The results in column 6 and 7 are for modification on vocal tract system characteristics using the LPCs, and these indicate similar trend as for duration modification. Larger values in column 6 and smaller values in column 7 show that the change is not effective perceptually.

In combination, pitch and other components seem to give significant improvement in perception of emotion in the synthesized utterance, as can be seen from columns 9, 11 and 15. Note that column 13 gives a lower score, as pitch is not modified in this case.

The results in column 16 are for the case where the LP residual is replaced by an LF model. In this case the scores are generally lower compared to the scores in column 15, where the LP residual of the neutral (source) utterance is retained in the synthesized speech. This indicates that including some residual information (although from neutral utterance) may be useful in perceiving the emotion in speech better.

The scores for subjective evaluation for experiments for Set 2 are given in Table 3. In this case the target is neutral utterance and the source is the emotion utterance. The objective is

to modify the emotion utterance by changing the components of speech selectively towards the neutral utterance. The somewhat larger scores in column 2 indicate that even though the pitch contour of the emotion utterance is modified to that of the neutral utterance, the effect of emotion is not reduced significantly. The lower scores in column 3 also support the above inference that the synthesized neutral, obtained by modifying the pitch contour of the emotion utterance, still retains the emotion characteristics of the source utterance. The large scores in column 4 clearly bring out the fact that duration modification cannot suppress the emotion characteristics of the source utterance.

Modification of only LPCs also gives scores (see column 6 & 7) similar (column 4 & 5) to duration modification. Modification of both pitch and duration improves the perception of characteristics of neutral utterance in the synthesized speech, as can be seen in column 9 in comparison with columns 3, 5 and 7. Similar observations can be made for modification in the combinations of pitch & LPCs and duration & LPCs, as shown in columns 11 and 13, respectively. When all the three components, namely, pitch, duration and LPCs are modified, then the synthesized neutral speech is much closer to the original (target) neutral utterance, as can be seen from column 15, when compared to column 9, 11 and 13. Also the scores are lower when the synthesized speech is compared with source (emotion) utterance, as shown in column 14.

It is interesting to note that in addition to the modification of pitch, duration and LPCs, if the LP residual of emotion utterance is replaced by an LF model, then the resulting similarity scores (see column 16) between original (neutral) utterance and the synthesized neutral speech are much higher, compared to the scores in column 15. This is an interesting result, which indicates that the LP residual component of the emotion (source) utterance seems to carry some perceptually useful information of emotion, which when replaced with an LF model, results in an utterance that is much more closer to the neutral (target) utterance. On the other hand, in Table 2 the absence of any LP residual brings down the scores in column 16 in comparison with the scores in column 15. Thus it appears that LP residual of neutral speech in some form seems to contribute to the perception of emotion in comparison with the case of total absence of it, as in the LF model case.

While in the above discussion general observations are made for all the emotions, there are significant differences in the scores for individual emotions. For example, for Set 1, the modification of pitch significantly improves the perception of 'happy' emotion in the synthesized speech, as shown in columns 2 and 3 in Table 2, and significantly reduces the perception of 'happy' emotion in the synthesized neutral speech, as shown in columns 2 and 3 in Table 3. The perception of 'disgust' emotion can be reduced significantly only when all the three components, namely, pitch, duration and LPCs are modified in the emotion utterance as can be seen from columns 2, 3, 14 and 15 in Table 3. Similar observations can be made for other emotions also.

6. Summary and conclusion

In this paper we have studied the significance of different components of speech in contributing to the perception of emotional state in a speaker's utterance. The two-way studies include transforming a neutral (source) utterance to emotion (target) utterance and an emotion (source) utterance to neutral (target) utterance by modifying one or more of the four components of speech. Subjective evaluations were carried out by comparing the synthesized speech with the target utterance to study the effectiveness of different components. It is interesting to note that, as expected, pitch contour plays significant role in the perception of emotion. Duration alone does not seem to contribute much. But duration and LPCs in combination with pitch contour seem to contribute to the perception of emotion in a significant way.

It is important to note that the above are general observations on the effectiveness of different components of speech for all emotions. There will be significant variations in the contribution of these components in an emotion-specific way. To draw meaningful conclusions on emotion-specific significance of these speech components, studies with more data (number of sentences) and with more speakers are needed.

In practice it is not possible to have the target utterance to determine the mapping of the components from source to target. Hence the challenge is to know what to modify in a given utterance to produce an utterance with desired emotion characteristics, when the reference (target) utterance is not available. It is also likely that all sound units in an utterance may not contribute to emotion. Different sets of units may contribute for different emotions, speakers and contexts. These are all challenging issues in the task of understanding how human beings produce and perceive the natural phenomenon of emotion.

7. Acknowledgements

This work is a part of ongoing research collaboration (2010-2014) on project Emotion between Speech and Vision Lab, IIIT, Hyderabad, India and SAIT, Samsung India Software Operations Pvt. Ltd., Bangalore, India. The authors would like to thank Prof. K. S. Rao for sharing the valuable database IITKGP SESC.

8. References

[1] Scherer, K. R., "Nonlinguistic vocal indicators of emotion and psychopathology", in *Emotions in personality and psychopathology*, C. E. Izard (Ed.), pp. 495-529, N.Y.: Plenum Press, 1979.

[2] M. Schroder and Martine Grice, "Expressing vocal effort in concatenative synthesis," in *Proc. 15th Int. Conf. Phonetic Sciences*, Barcelona, Spain, pp. 2589-2592, 2003.

[3] K. Hirose, K. Sato, Y. Asano and N. Minematsu, "Synthesis of F0 contours using generation process model parameters predicted from unlabeled corpora: application to emotional speech synthesis," *Speech Communication*, vol. 46, no. 3-4, pp. 385-404, 2005.

[4] S. G. Koolagudi, S. Maity, V. A. Kumar, S. Chakravarti and K. S. Rao, "IITKGP-SESC: Speech database for emotion analysis", *Communications in Computer and Information Science*, JIIT University, Noida, India: Springer, ISSN-1865-0929, Aug. 2009.

[5] P. Gangamohan, V. K. Mittal and B. Yegnanarayana, "A Flexible Analysis Synthesis Tool (FAST) for studying the characteristic features of emotion in speech", in *Proc. 9th IEEE Consumer Communications and Networking Conference*, Jan. 2012.

[6] H. Sakoe and S. Chila, "Dynamic programming algorithm and optimisation for spoken word Recognition", *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 26, pp 43-49, 1978.

[7] K. S. R. Murty and B. Yegnanarayana, "Epoch extraction from speech signals", *IEEE Trans. Acoust., Speech, Lang. Process.*, vol. 16, no. 8, pp. 1602-1614, Nov. 2008.

[8] K. S. Rao and B. Yegnanarayana, "Prosody modification using instants of significant excitation", *IEEE Trans. Acoust., Speech, Lang. Process.*, vol. 14, pp. 972-980, May 2006.

[9] G. Fant, J. Liljencrants and Q. Lin,"A four-parameter model of glottal flow", *Quarterly Progress Status Report, Speech Trans. Lab., KTH-Sweden*, vol. 26, no. 4, pp. 001-013, 1985.