Investigating the role of phoneme-level modifications in emotional speech resynthesis

Murtaza Bulut, Carlos Busso, Serdar Yildirim, Abe Kazemzadeh, Chul Min Lee,
Sungbok Lee, Shrikanth Narayanan

Department of Electrical Engineering
Viterbi School of Engineering, University of Southern California, Los Angeles, CA, USA
Speech Analysis and Interpretation Laboratory, http://sail.usc.edu
mbulut@usc.edu

Abstract

Recent studies in our lab show that emotions in speech are manifested as, besides supra-segmental trends, distinct variations in phoneme-level prosodic and spectral parameters. In this paper, we further investigate the significance of this finding in the context of emotional speech synthesis. Specifically, we study phoneme-level signal property manipulation in transforming the emotional information conveyed in a speech utterance. We analyze the effect of individual and combined modifications of F0, duration, energy and spectrum using data recorded by a professional actress with happy, angry, sad and neutral expressiveness. We use content matched source-target pairs and apply TD-PSOLA for prosody and LPC for spectrum modifications by directly extracting the required parameters from the target speech. Listening tests conducted with 10 naive raters show that modification of prosody and spectral envelope parameters by themselves is not sufficient. However, when applied together, modifying spectrum and prosody at the phone level gives successful results for most emotion pairs, except conversion to happy targets. We also observe that at the phoneme level, spectral envelope modifications are more effective than local prosodic modifications; and that, duration modifications are more effective than pitch modifications. The results confirm our hypothesis that phoneme level modifications can be used to fine tune the ensuing suprasegmental-parameter-based modifications to improve the overall quality of synthesized emotions.

1. Introduction

Emotion resynthesis (or conversion) is an adaptation technique where the input emotional speech is modified so that the output speech is perceived as conveying a new emotion. The parameters of the input speech emotion are adapted to the target emotion and then the final output is resynthesized using the new parameters.

Emotion conversion is a novel research area which resembles voice conversion (VC) in terms of the underlying techniques, the most important distinction being that the effect of the prosody can not be ignored or understudied (because the speech emotion and prosody are strongly tied to each other [1, 2]), as it is usually done in the conventional VC algorithms where pitch modifications are generally utilized only by matching the converted average pitch to the average target pitch. Applying segmental pitch modifications [3], copying the target pitch contours [4], or using heterogeneous training vectors including both spectral coefficients and normalized pitch [5] has shown that incorporation of prosody modifications improves the VC results.

Motivated by the fact that distinct emotional coloring is also present at the phoneme level (especially for back and low vowels), as shown in our own recent emotional speech analysis studies [6, 7], in this paper we investigate the applicability of segmental modification of duration, pitch, energy and spectral envelope parameters for the resynthesis of four emotions, happy, angry, sad and neutral. We study the effect of each parameter on emotion perception, when applied on various source-target emotional pairs using TD-PSOLA [8] and LPC synthesis methods, to show that by modifying phonemes' acoustic features we can change the emotional content of the whole sentence. Although here we tested only at phoneme-level, such segmental modifications can be extended to diphones, syllables and words, and they can be integrated in concatenative speech synthesizers as pre-processing of concatenation units before synthesis, with the purpose of improving the synthesized emotion's quality for the whole sentence.

For data collection we recorded sentences uttered by a professional actress who was instructed to produce four full-blown emotions, anger, happiness sadness and neutral, for the sentences of identical content. The validity of whether the recorded data can be correctly identified in terms of its emotional content was validated by conducting listening tests with 10 naive raters. The same approach was followed for the evaluation of the final resynthesized output utterances as well. Considering the fact that the interpretation of emotions can sometimes differ due to personal, cultural and many other experiences, it is not unusual that the raters will disagree on the emotional content of a presented utterance. As pointed out in many studies, the boundaries between emotions are fuzzy and during the evaluation this fact should be taken into account [9, 10]. A good review of these emotion research issues can be found in [1].

In the rest of the paper, we first describe the used dataset and then introduce the proposed conversion system in section 3. Evaluation results are presented in section 4 and discussed in section 5.

2. Dataset description

The sentences that we used for this study are (1) *"This is some union we've got"* and (2) *"The store closes at twelve"*. Formant, pitch and energy plots for these two sentences are presented in Figure 1 and 2. In terms of the pitch contour and energy plots (Figure 2) we see that anger is somewhat similar to happiness and sadness is similar to neutral. The plots of the first two formants (Figure 1) for the vowels show that the vowel formants vary based on sentence emotions.

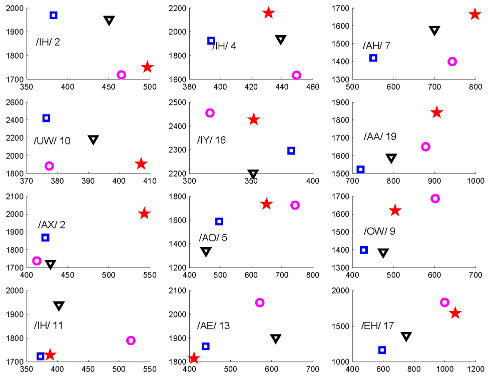

Figure 1: *F1-F2 plots for sentence vowels. Top 6 plots are for sentence 1 vowels. Numbers indicate the position of the vowel in the sentence. Happy is represented by $\star$, angry is o, sad is $\Box$, neutral is ∇.*

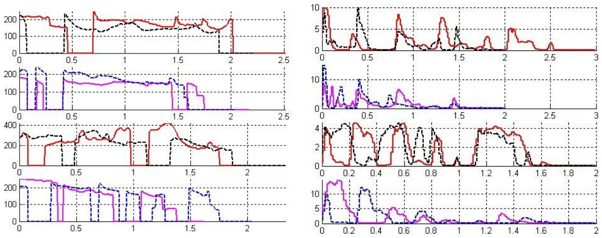

Figure 2: *Pitch contour (left) and energy contour of the sentences. Top two plots are for sentence 1 and first of these top plots is the plot of happy and angry(dashed) and the second plot is of sad and neutral(dashed) emotions.*

3. System description

A schematic diagram of our emotion conversion system is shown in Figure 3.

All of the modifications are performed on the source signal ($s[n]$), based on the features extracted from the target signal ($t[n]$). For pre-processing, the source and target speech are normalized to the same intensity level (i.e., 70db) and the label boundaries for each utterance are manually extracted. Next, we calculate the pitchmarks of both signals and use the target pitchmarks to modify the pitch contour and duration of the source by applying the TD-PSOLA [8] algorithm. The energy modification is performed next, by normalizing the power of the TD-PSOLA output, so that it will match the target signal's power (i.e., $Output = Output * (targetpower/outputpower)^{0.5}$). The new output, $s_1[n]$, is the prosody-modified version of the source. For spectral envelope modification, we calculate the linear prediction coefficients (LPC) using a prediction error filter ($A(z)$) of order 16 and Hanning-windowed 20ms long frames interlaced with 50% overlap. The pre-emphasis coefficient was set to 0.9. To change the spectral characteristics of the source, we filter the source residual ($e_s[n]$) with the inverse of the target error filter to get $s_2[n]$, which is then futher prosody modified, as described above, to produce the final result, $s_{12}[n]$. Alignment of source and target signals is performed automatically, by adding or deleting frames from the mid-regions (where the spectrum is relatively stable) of the processed segments. While all modification are performed for voiced phonemes, only duration and energy are changed for the rest.

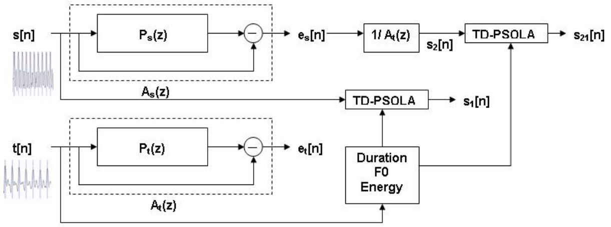

Figure 3: *LPC and TD-PSOLA based emotion conversion system. $s[n]$ and $t[n]$ are the source and target, respectively. $A(z)$ indicates the inverse filter used for the calculation of the residual $e[n]$. $s_1[n]$, $s_2[n]$, and $s_{21}[n]$ are the outputs obtained by modifying the input signal using only TD-PSOLA, only LPC synthesis, and both LPC and TD-PSOLA, respectively.*

3.1. Test stimuli

The proposed conversion was applied on the two target test sentences. For each of the sentences there were 12 source-target pairs. They were: (1) happy-angry, (2) happy-sad, (3) happy-neutral, (4) angry-happy, (5) angry-sad, (6) angry-neutral, (7) sad-happy, (8) sad-angry, (9) sad-neutral, (10) neutral-happy, (11) neutral-angry, (12) neutral-sad. In order to observe the effect of the individual parameter changes, each of these pairs was modified in a controlled manner by changing one parameter at a time. The list of the feature modifications we investigated include: (1) only pitch, (2) pitch and energy, (3) only duration, (4) duration and energy, (5) duration and pitch, (6) duration, pitch and energy, (7) only spectrum, (8) pitch and spectrum, (9) pitch, energy and spectrum, (10) duration and spectrum, (11) duration, energy and spectrum, (12) duration, pitch and spectrum (13) duration, pitch, energy and spectrum. In total we synthesized 156 (12x13) sentences for each test sentence. Together with 312 (2x156) synthetic signals and with the inclusion of the original sentences (2 happy, 2 angry, 2 sad and 2 neutral), our test set consisted of 320 stimuli.

4. Evaluation

In this section we describe the listening test set up and present the evaluation results.

4.1. Listening experiment

Assessment of the output emotion categories is achieved by conducting subjective listening tests with naive listeners. Ten listeners participated in the experiment and each of the listeners was presented with 320 sentences, which consisted of the results of all modifications as well as the original utterances. The test stimuli were presented in random order in order to eliminate any correlative effects in decision making. Headphones were used and the users were given the freedom to adjust the volume and to listen to the current sentence as many times they wanted, however once done they were not given the opportunity to return back. The test was organized as a forced-choice experiment, where the raters were required to decide on one of the following five choices: (1) happy, (2) angry, (3) sad, (4) neutral and (5) other. The inclusion of *other* option was to provide a way for including intermediate (i.e, fuzzy) emotional categories that can happen as a result of the modifications. The average test duration was 25 minutes per listener.

4.2. Listening test results

Results for the original (unmodified) and synthesized utterances are presented in Table 1 and Table 2, respectively.

4.2.1. Original sentences

Listening test results for the original sentences, displayed in table 1, show that our speaker successfully elicits all of the emotions. The kappa statistics, $\kappa = 0.70$, $\alpha < 0.01$, show strong rater agreement. These results serve as an upper bound.

	H1	A1	S1	N1	H2	A2	S2	N2
Happy	70	20	0	10	100	0	0	0
Angry	10	80	0	10	0	100	0	0
Sad	0	10	70	20	0	0	80	20
Neutral	0	0	0	100	0	0	0	100

Table 1: *Listening test results for the two original sentences ($\kappa = 0.70$, $\alpha < 0.01$). Emotion categories are listed in the first column and the results are presented in percentages.*

4.2.2. Resynthesized sentences

The two best results for each possible pair are shown in Table 2. For full table of results please check the website [1]. Kappa statistics calculated for all 312 rater responses are as follows: Sentence 1:$\kappa = 0.25$,$\alpha < 0.01$, Sentence 2:$\kappa = 0.36$,$\alpha < 0.01$. These values are much lower compared to the original sentences, and they indicate the inherent difficulties in evaluating emotional [9] and synthetic speech. In Table 2, the results for sentence 1 and 2 are presented separately. The number shown in the parenthesis indicates the performed modification, as described in section 3.1. The results are in 10% multiples, for instance, for happy-to-sad conversion (h2s, i.e. source is happy, target is sad emotion) of sentence 1, when method (9) was applied, the final resynthesized output was recognized as 20% *happy*, 60% *sad*, 10% *neutral* and 10% as *other* (not shown in the table). The results indicate successful conversion, except for the following pairs: Sent.2-a2h, s2h, n2h. Apparently, synthesis of happiness is not achieved within an accepteble level, emphasizing the fact, as shown in many other papers, that special attention should be paid to it, because the signal distortions arising due to the modifications influence listener's judgments especially of happy emotion.

In addition, we note that happy emotion was usually confused with angry emotion. This confusion can be related to the similarity of the acoustic features of happy and angry sentences as shown in Figure 2. Kendall's tau-b correlation calculation (for all 312 sentences) show positive correlation between numbers of *happy* and *angry* responses ($\tau_b = 0.074$, $\alpha < 0.108$). Although, we see that sad and neutral sentences also have resembling acoustic features, the correlation between neutral-sad responses was insignificant ($\tau_b = 0.018$, $\alpha < 0.683$). For all other pairs significant negative correlation was observed (angry-neutral: $\tau_b = -0.433$, $\alpha < 0.01$; angry-sad: $\tau_b = -0.424$, $\alpha < 0.01$; angry-other: $\tau_b = -0.213$, $\alpha < 0.01$; happy-neutral: $\tau_b = -0.290$, $\alpha < 0.01$; happy-sad: $\tau_b = -0.308$, $\alpha < 0.01$; happy-other: $\tau_b = -0.107$, $\alpha < 0.025$). It is also interesting to note the positive correlation between *neutral-other* responses ($\tau_b = 0.276$,$\alpha < 0.01$), indicating that raters tend to choose *neutral* whenever the appropriate emotion choice is not listed in the evaluation test.

5. Discussion

In this study we copy all of the information from the target, and thus our approach should not be directly compared to the unsupervised and automated processes. Here, we aim to provide insights into the usefulness and feasibility of phoneme-level modifications for emotion control, with a goal toward creating fully automated conversion rules in the future. This section details the effects of individual and combined application of prosody and spectral modifications on the emotions.

5.1. Prosody modifications

Prosody parameters -duration, pitch and energy- at the supra-segmental level are considered the defining factor for emotions [1, 2, 11]. However, to make concatenative synthesis systems more useful and flexible we need to be able to modify segmental-level properties as well. Our analysis results have reported distinct emotion effects at the segmental level speech features [6], which suggest that by performing segmental modifications we can alter the emotion of the whole sentence. The results (available on the website[1]) show that when applied without any spectral envelope modifications, prosody parameter modifications done locally at the phoneme-level are not effective in changing the emotion perception. This is because, it is not possible to exactly match and reproduce the target sentence prosody only by doing local prosody modifications.

Local modifications are not sufficient due to two main reasons: (1) speech prosody features, such as inter-word silences, fluent pauses, hesitation pauses, stress pattern and shimmer [12], that we did not modify here, have significant effect on the sentence prosody; (2) TD-PSOLA modifications at phoneme level, especially when the ratio between source and target phone parameters is larger (smaller) than 2 (0.5), introduce perceptible and visible artifacts, which must be smoothed by further sentence level processing. Thus, the evaluation results support our hypothesis that for the local prosody modifications to be effective and meaningful they should always be followed by additional supra-segmental level modifications.

5.2. Spectrum modifications

As outlined in the introduction, voice personality can be successfully changed by transforming frequency parameters [3, 4, 5]. Tenseness, creakiness, laxness, breathiness which are all attributes of the voice quality can be also changed by spectral modifications. These attributes are closely related to the emotional content of speech [13], so one can expect that changing them will be useful for emotion re-synthesis. We test this by directly using the target LPCs to modify source speech.

The complete table of results (available on the website) show that spectrum modifications increase the ambiguity between the emotion classes (i.e., emotions are more confused with one another due to the new emotional coloring) but not to any significant level (except for Sent 2: a2s, n2s) to cause change in the emotion category. We also observe that for phonemes, spectral modifications are more effective than local prosody modifications in shifting the source emotion (in the emotional space) closer to the target emotion. This justifies our hypothesis that transforming the phonemes' spectral properties

[1] http://sail.usc.edu/∼mbulut/euro05.html

	h2a	h2s	h2n	a2h	a2s	a2n
Sent. 1	(11) h:4 a:4 s:0 n:1 (13) h:6 a:4 s:0 n:0	(9) h:2 a:0 s:6 n:1 (12) h:1 a:0 s:5 n:1	(11) h:0 a:2 s:1 n:6 (13) h:0 a:1 s:1 n:6	(7) h:3 a:4 s:1 n:1 (11) h:4 a:2 s:0 n:3	(8) h:0 a:3 s:6 n:0 (12) h:0 a:2 s:4 n:2	(12) h:0 a:2 s:2 n:5 (13) h:0 a:2 s:1 n:6
Sent. 2	(11) h:3 a:6 s:0 n:1 (13) h:4 a:5 s:1 n:0	(12) h:1 a:0 s:3 n:4 (13) h:3 a:0 s:3 n:2	(10) h:5 a:0 s:0 n:3 (11) h:4 a:0 s:2 n:3	(12) h:1 a:8 s:0 n:1 (13) h:1 a:9 s:0 n:0	(7) h:0 a:1 s:5 n:1 (9) h:0 a:1 s:5 n:0	(10) h:0 a:4 s:2 n:3 (13) h:0 a:3 s:1 n:5

	s2h	s2a	s2n	n2h	n2a	n2s
Sent. 1	(8) h:3 a:2 s:0 n:3 (13) h:2 a:1 s:3 n:3	(9) h:0 a:4 s:2 n:2 (13) h:1 a:5 s:0 n:3	(12) h:0 a:0 s:2 n:7 (13) h:0 a:0 s:1 n:8	(11) h:1 a:3 s:1 n:4 (12) h:1 a:2 s:3 n:2	(12) h:1 a:4 s:0 n:3 (13) h:1 a:5 s:0 n:2	(10) h:0 a:0 s:5 n:3 (12) h:1 a:0 s:5 n:3
Sent. 2	(12) h:0 a:4 s:3 n:1 (13) h:3 a:1 s:4 n:1	(9) h:0 a:5 s:1 n:0 (13) h:1 a:6 s:0 n:0	(11) h:0 a:0 s:3 n:6 (12) h:0 a:0 s:3 n:6	(12) h:0 a:4 s:3 n:1 (13) h:3 a:3 s:2 n:0	(11) h:0 a:5 s:1 n:2 (13) h:2 a:6 s:0 n:1	(7) h:0 a:0 s:8 n:2 (8) h:0 a:0 s:4 n:5

Table 2: *Listening test results (in 10% multiples) for some selected modifications. h,a,s,n indicate happy, angry, sad and neutral, respectively. The numbers in paranthesis refer to the modification type as explained in section 3.1. h2a means that source is happy and target is angry.*

can be effectively used in speech synthesizers to add new emotional content to the synthesized speech.

5.3. Prosody and spectrum modification combination

Concurrently applied prosodic and spectral modifications give better results than their individual applications. As expected, when all 4 variables (spectrum, duration, pitch and energy) are modified the results improve. From Table 2 we see that the best results are achieved for following pairs, Sent 1: h2s, h2n, a2s, a2n, s2n; Sent 2: a2s, s2a, s2n, n2a, n2s.

Our evaluation experiments do not show a particular trend in the results that can be associated with each parameter individually. For example, starting with only spectrally modified sentences, when we changed the pitch the recognition rates for some pairs (Sent1: h2a, h2s, a2s, a2n, s2h, n2s; Sent2: a2s, s2a, n2s) improved while for the others there was no improvement. The same observation is also valid when only duration modifications were applied to the spectrum-only modified sentences (pairs that improve are sent1: h2a, h2s, a2s, a2n, s2a, s2n, n2h, n2a, n2s; sent2: h2a, h2n, a2n, s2a, n2h, n2a). Inclusion of the modification of an additional prosodic feature (such as duration, pitch and energy), generally speaking, improved the results. In addition, duration changes proved to be more effective than the pitch changes, which in turn were more effective than the energy changes when applied on spectrally modified phonemes.

The results indicate following trend, in terms of producing successful conversion, among proposed feature modification methods: (13) duration, energy, pitch, spectrum $>$ (12) duration, pitch, spectrum $>$ (11) duration, energy, spectrum $>$ (9) pitch, energy, spectrum $>$ (10) duration, spectrum $>$ (8) pitch, spectrum $>$ (7) spectrum.

6. Conclusion

In this paper we studied the conversion of one emotion into another by modifying prosody and spectral envelope at phoneme-level. Our results show that individually modified local prosody and spectrum add new emotional coloring to the source emotion. However, they are not sufficient by themselves to elicit the target emotion. Comparing the two, we see that at phoneme-level, spectral envelope modifications are more effective than local prosodic modifications, and for local prosody, duration modifications are more effective than pitch modifications. When applied together, local prosody and spectrum modifications successfully transformed the emotion of the source speech to the target emotion.

These results support our hypothesis that combining phoneme level and supra-segmental level modifications can be a useful framework to model the emotional content of synthesized speech. Furthermore, for an emotion synthesizer to be fully successful additional linguistic layers -from phonemic and lexical context to syntactic and discourse structures- should be carefully accounted for. Much of those details are still unknown and are a subject of ongoing work.

7. References

[1] Scherer, K.R., ”Vocal communication of emotion: A review of research paradigms”, Speech Communication, vol. 40 (1-2), pg. 227-256, 2003.

[2] Cowie, R., Cowie, E.D., Tsapatsoulis, N., Votsis, G., Kollias, S., Fellenz, W., and Taylor, J., ”Emotion recognition in human-computer interaction”, *IEEE Sig. Proc. Mag.*, vol. 18(1), pp. 32 80, Jan 2001.

[3] Turk, O., Arslan, L.,M.,”Voice conversion methods for vocal tract and pitch contour modification”, *EUROSPEECH-2003*, 2845-2848, 2003.

[4] Valbret, H., Moulines, E., and Tubach, J.P., ”Voice Transformation Using PSOLA Technique”, *Speech Communication*, 11, 175–187, 1992.

[5] Najjary, T., Rosec, O., Chonavel, T., ”A voice conversion method based on joint pitch and spectral envelope transformation”, *ICSLP*, 2004.

[6] Yildirim, S., Bulut, M., Lee, C.M., Kazemzadeh, A., Busso, C., Deng, Z., Lee, S., and Narayanan, S., ”An acoustic study of emotions in expressed speech”, *ICSLP*, 2004.

[7] Lee, C.M., Yildirim, S., Bulut, M., Kazemzadeh, A., Busso, C., Deng, Z., Lee, S., and Narayanan, S., ”Emotion recognition based on phoneme classes”, *ICSLP*, 2004.

[8] Moulines, E., and Charpentier, F., ”Pitch-Synchronous Waveform Processing Techniques for Text-to-Speech Synthesis Using Diphones”, *Speech Communication*, vol. 9, pp. 453-467, December 1990.

[9] Steidl, S., Levit, M., Batliner, A, Nth, E, Niemann, H., ”Of all Things the Measure is Man. Automatic Classification of Emotions and Inter-labeller Consistency”, *ICASSP*, 2005.

[10] Campbell, N., ”Perception of Affect in Speech - towards an Automatic Processing of Paralinguistic Information in Spoken Conversation”, *ICSLP*, 2004.

[11] Bulut, M., Narayanan, S., and Syrdal, A., “Expressive Speech Synthesis Using a Concatenative Synthesizer”, *Proc. of ICSLP*, Denver, CO, 2002.

[12] Cahn, J.E., ”Generating Expressions in Synthesized Speech”, Master’s Thesis, MIT, 1989. http://www.media.mit.edu/∼cahn/masters-thesis.html

[13] Gobl, C. and N Chasaide, A., ”The role of voice quality in communicating emotion, mood and attitude”, *Speech Communication*, vol. 40. pp. 189-212, 2003.