A STATISTICAL APPROACH FOR MODELING PROSODY FEATURES USING POS TAGS FOR EMOTIONAL SPEECH SYNTHESIS

Murtaza Bulut, Sungbok Lee and Shrikanth Narayanan**

University of Southern California, Los Angeles, CA
Department of Electrical Engineering, *Also, Department of Linguistics
mbulut@usc.edu, sungbokl@usc.edu, shri@sipi.usc.edu

ABSTRACT

Deriving statistical models for emotional speech processing is a challenging problem because of the highly varying nature of emotion expressions. We address this problem by modeling prosodic parameter differences at the part of speech (POS) level for emotional utterances for the purpose of emotional speech synthesis. Synthesis at the POS level is appealing because POS tags carry salient information conveying speech prominence. Analysis of energy, duration and F0 differences between matching neutral-angry, neutral-sad and neutral-happy emotional utterance pairs shows that Gaussian distributions can be used to model the parameter differences. Pairwise comparisons of POS features reveal that it is more probable that the normalized mean and median energy of sad POS tags are larger than neutral, angry or happy POS tags. They also show that for particular tags it is more likely that angry emotion has higher F0 median than happy emotion, and that sad emotion has higher F0 median than neutral emotion. Experiments of conversion of neutral speech into emotional speech using the Gaussian probability functions provide helpful insights into the application of statistical models in speech synthesis.

Index Terms— POS, emotion, prosody, energy, conversion

1. INTRODUCTION

Synthesis of emotional speech is a complex process which requires different features and segments to be modified in accord with each other. The important parameters which have been attributed to emotional effect generation are F0, duration and energy [1]. Importantly, it has been shown that emotional modulations in speech production interplay with the goals of achieving linguistic contrasts [2]. As a result, any speech synthesis or conversion scheme for emotional effects is inherently constrained by the underlying linguistic structure of an utterance and hence should be accounted for as a part of the process. In prior work, we explored emotional modifications at the phoneme level [3]. In this paper we analyze the prosody parameters at the level of part of speech (POS) categories, and present a probabilistic approach of how they can be applied to emotional speech synthesis through conversion of neutral speech. Motivation for working in POS level comes from the potential of higher quality synthesized speech than phoneme level modifications.

In the presented approach, instead of modeling the characteristics of individual emotional categories, we propose modeling differences between emotions in POS level. The results show that these differences can be parameterized by fitting into Gaussian distributions.

Part of speech tags are popularly used in natural language processing and in many text-to-speech systems. In TTS systems, typically instead of considering the whole set of possible POS tags, the focus has been mainly on content/function word distinction [4]. For example in the Affect Editor [5] the content/function word distinction was used in adjusting the ”stress frequency” and ”hesitation pauses” parameters. Another example is HMM based synthesis [6], where POS category was one of the contextual factors used for training of limited speech database emotional synthesizer. This paper looks at prosodic parameter differences at the POS level for emotional utterances.

Duration and F0 parameters have been generally paid more attention than energy in the analyses of emotional speech [1]. In this work we concentrate on investigating energy and show the differences between emotions with respect to POS energy maximum, median and mean values. The idea for modeling energy at the POS level is based on the reports showing that energy is an essential component in prominence perception [7] and that speech prominence scores are related to POS categories [8]. It can be expected that inclusion of emotion specific energy modifications together with F0 and duration modifications will improve emotional speech synthesis results.

2. PART OF SPEECH LABELING AND FEATURE EXTRACTION

Part of speech tagging of the sentences was performed using the Charniak POS parser [9]. This parser is based on a probabilistic generative model and it achieves 90.1% average precision/recall performance on sentences shorter than 40 words. For a given sentence input it generates Penn tree-bank style parse tress [10]. For an example input sentence, “Count the number of teaspoons of soysauce that you add” (a sentence from the TIMIT database), the output generated by the parser is: “(S1 (FRAG (-LRB- -LRB-) (“ “) (S (VP (VB count) (NP (NP (DT the) (NN number)) (PP (IN of) (NP (NP (NNS teaspoons)) (PP (IN of) (NP (NN soysauce) (SBAR (IN that) (S (NP (PRP you)) (VP (VBP add)))))))))) (. .) (” ”) (-RRB- -RRB-)))”. From the generated output one can easily identify the Verb Phrase (VP), Noun Phrase (NP), etc., boundaries and POS tags for individual words.

The time domain word boundaries for each utterance were automatically estimated using adapted-HMM models trained (using HTK toolkit) on the TIMIT database and adapted by maximum likelihood linear regression model using our emotional speech data.

Utterance F0 contours were measured using Praat software. After smoothing the F0 vector by a median filter of length 3, the mean, median, maximum, minimum statistics were calculated for each word.

Average magnitude function was used for energy contour calculations. In this method, instead of the squares of individual values, their absolute values are summed over a shifting short-time win-

dow [11]. In our case, a Hamming window of 240 samples (0.015 seconds) was used. The average magnitude function was preferred over the standard (RMS) energy calculations because of its smaller dynamic range.

Two type of normalizations were performed during energy calculations. First, the digitized speech file was normalized so that it has a maximum sample amplitude of 1. Then, the utterance energy contour was calculated for the normalized waveform. The second normalization was performed on the calculated energy contour by dividing the energy vector by its maximum. These normalizations ensure that effects of voice level amplitude variations due to different emotion expressions are minimized and that energy contours of different utterances (and different speakers) can be directly compared.

It is well known that angry and happy speech possess greater energy than neutral speech and that sad speech usually has the lowest energy [1]. After the normalizations, emotion dependent differences in voice level are not observable anymore. Instead, new energy contour characteristics such as the relative energies of each word (to each other) become visible. Having the relative energy information between words (i.e., ratio of energies) enables us to study how the shape of the energy contour is varied from emotion to emotion.

3. COMPARATIVE ANALYSIS OF PROSODY AT POS LEVEL IN EMOTIONAL UTTERANCES

The speech material analyzed in this paper consists of 249, 201 and 192 sentences recorded by two female speakers and a male speaker, respectively. All of these semantically neutral sentences were uttered (i.e., acted) in 4 different ways, namely, expressing anger, happiness, sadness and with neutral state, resulting in a total of 2568 (= 4 x 642) utterances.

Having lexically matching utterances expressed in different emotions, as in our case, enables us to directly compare the acoustic features for different emotions. For every POS tag and speaker, we analyzed the differences in acoustic features for each one of the neutral-angry, neutral-happy, neutral-sad, angry-happy, angry-sad and happy-sad emotional pairs. The goal was to learn patterns to inform design of statistical emotion conversion schemes. The specific focus of our analysis was on the POS tags of content words as they are known to affect the emotional and perceptual quality the most [4, 5]. These particular POS tags were, JJ (representing adjectives, 413 (shows the number of occurrences in the analyzed emotional dataset)), NN (singular or mass nouns, 804), NNP (proper singular nouns, 124), NNS (plural nouns, 214), PRP (personal pronouns, 472), PRP$ (possessive pronouns, 136), RB (adverbs, 355), VB (base form verbs, 261), VBD (past tense verbs, 112), VBG (gerund or present particle verbs, 132), VBN (past particle verbs, 114), VBP (non-3rd person singular present verbs, 84) and VBZ (3rd person singular present verbs, 57).

The results (averaged over all speakers) comparing the energy, duration and F0 feature values across emotions for the content POS tags are shown in Fig. 1. Shown in the figure is the probability that for an emotional pair (emotion1-emotion2), the second emotion (emotion 2) has higher feature values than the first emotion (emotion 1). The probabilities were computed by counting the number of the instances where the target event occurs (i.e., emotion 2 feature $>$ emotion 1 feature) between matched utterances and dividing the total count by the total number of the compared utterances. For analysis we define the $0.4 < p < 0.6$ region as an equal chance region and specifically concentrate on POS tags falling outside this region.

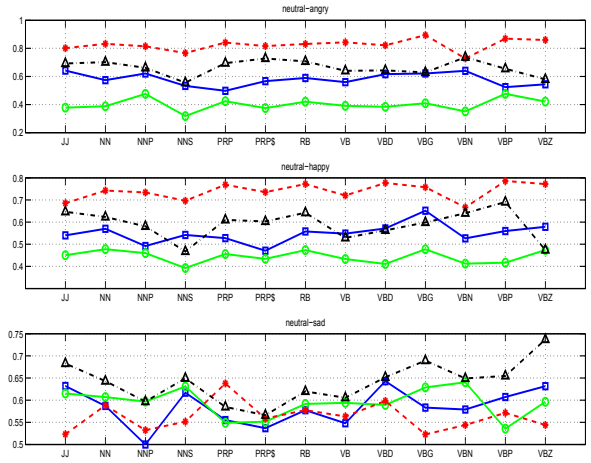


Fig. 1. Probabilities of emotion 2 having greater (acoustic feature) value than emotion 1. The displayed emotional pairs (emotion 1 - emotion 2) are neutral-angry (top panel), neutral-happy (middle panel) and neutral-sad (bottom panel). The plotted lines are for energy maximum (□), energy median (○), pitch median (∗) and duration (△).

Analyzing the results for the maximum energy of words (i.e., the maximum value of the energy contour falling inside a word boundary) for neutral-angry pairs, it is seen that in most utterances angry words have higher maximum energy than neutral words. Especially, this difference was observable for JJ, NNP, VBD, VBG and VBN tags, for which the probabilities that an angry word will have higher maximum energy than the matching neutral word were higher than 0.6. In contrast, this behavior was true only for VBG in the neutral-happy matching pairs. For other tags the probabilities were also generally higher than 0.5 but they fell inside the defined equal chance region (i.e., $0.4 < p < 0.6$). Interesting results were observed for neutral-sad pairs. In general, without any normalizations sad words have less energy than neutral words. However, this behavior reversed when the energy contour was normalized (as it is in this case). In this particular dataset, it was calculated that sad JJ, NNS, VBD, VBP, VBZ words have higher maximum energy than their neutral counterparts in more than 60% of the cases. These results indicate shift of utterance level energy maximum to different tags as emotion varies. They also show the POS tags that need to be paid more attention for each emotion.

ANOVA (and similarly Kruskal-Wallis) analysis showed significant differences between the maximum values of the content POS tags (listed above) for different emotions ($F = 36.61$, $p < 0.05$; $\chi^2 = 100.37$, $p < 0.05$). Multiple comparisons of means and medians (performed using Matlab statistical toolbox) showed that neutral maximum values have the lowest mean (median) ($neutral < happy < sad < angry$), which was significantly different than the means (and medians) of the maximums of the emotional speech. The results also show that angry maximums were significantly different from happy maximums, but not from the sad maximums.

In order to observe how energy is distributed among words in an utterance depending on the emotion, absolute differences between sorted POS tag energy maximums in each utterance were compared. The mean of differences between the maximums of the two POS tags with the highest maximum energies were 0.1678, 0.1851, 0.2085 and 0.2183 (and significantly different based on ANOVA analysis, $F = 14.24$, $p < 0.05$) for angry, happy, sad and neutral emo-

tions, respectively. Pairwise comparisons showed that neutral or sad speech is significantly different than happy or angry speech. The differences between neutral-sad and happy-angry pairs were not significant. When the three or four POS tags with the highest maximums were considered, the mean of differences of maximums were again significantly different ($F = 26.88$, $p < 0.05$; $F = 22.99$, $p < 0.05$). For these cases, pairwise comparisons showed that neutral speech is significantly different than emotional (i.e., angry, happy or sad) speech and that angry speech is significantly different than happy or sad speech. There were not any significant differences between happy and sad speech. When the differences between the maximums of the five tags with highest maximums were considered the differences (in terms of statistical significance) between happy, angry and sad speech were not observable anymore. However, neutral speech was still significantly different than these emotions. When all words in the sentences were considered there were not any significant differences between emotional and neutral speech in terms of the mean of the differences of sorted word maximum energy values.

Conventional energy calculations showed that there is a difference between energy levels of emotional utterances. These new results indicate that in neutral utterances the differences between the energy maximums of words is much more salient than in emotional speech. In contrast, in emotional utterances it is more common to see energy contour peaking at multiple word locations. The results show that there are significant differences in how the energy is distributed among words in emotional utterances. These observations are especially beneficial for modeling the shape of utterance energy contours for various emotions.

Comparisons of energy median values of POS tags, which were calculated from the normalized utterance energy contours also reveal interesting results. They show that energy in neutral and especially in sad words is spread over the whole word in contrast to angry words where the energy is concentrated in particular locations.

These conclusions are derived from the results for energy median (and also energy mean), where we see that in more than 60% of instances angry JJ, NN, NNS, PRP$, VB, VBD, VBN words had smaller median values than their neutral counterparts. Sad JJ, NN, NNS, VBG, VBN words had higher median energies than their neutral pairs in more that 60% of the instances. Similarly, sad JJ, NN, NNS, RB, VB, VBD, VBN, VBZ words were more likely to have higher medians than their happy or angry pairs.

ANOVA and multiple comparisons tests were performed on the median energy values of the analyzed content POS tags. The results show significant differences ($F = 44.45$, $p < 0.05$) between the four emotions. Multiple comparisons showed that differences between neutral and happy emotions were not significant, as other pairwise differences were significant. Similar results were calculated for the POS energy mean comparison, where the ANOVA statistics was $F = 46.07$, $p < 0.05$.

Pairwise word duration comparisons show that in most cases sad words have longer and neutral words have shorter durations. It was also interesting to observe that in most cases (with probability $p > 0.5$) happy words were shorter than angry words. ANOVA analysis for durations showed significant differences between emotions ($F = 28.95$, $p < 0.05$).

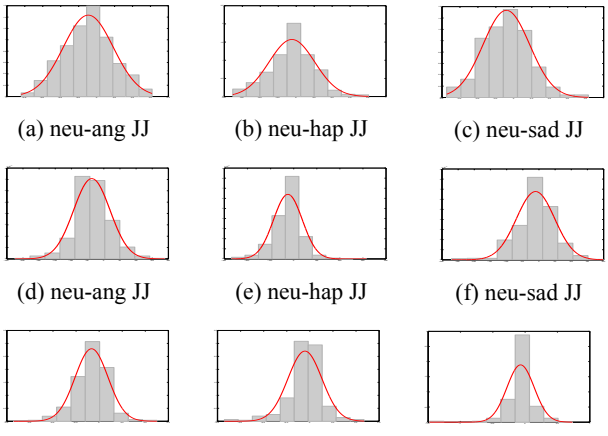
For F0 median (and also mean) the results are as expected. Angry or happy words have higher F0 median (mean) values than the neutral or sad words. The interesting and important results here are for the angry-happy word pairs. It is observed that the probability that an angry word (JJ, NNP, NNS, RB, VB, VBD, VBG, VBN, VBP, VBZ) will have higher F0 median (mean) than its happy pair is greater than 0.6. Considering that in many emotional speech studies the F0 values of happy and angry emotions were difficult to differentiate, these results are very helpful.

The probabilities calculated for neutral-sad words were mostly inside the equal chance region, except for F0 median of PRP words and F0 mean of NNP, PRP, RB and VBD words, for which it was more probable ($p > 0.6$) that sad words had higher values. Such differences may be important in emotional speech synthesis. The differences in the F0 median (mean) values were significant, ANOVA results were $F = 261.08$, $p < 0.05$ ($F = 215.61$, $p < 0.05$).

4. STATISTICAL MODELING OF EMOTIONAL DIFFERENCES

Modeling the acoustic features of speech at the word level using POS categories is an attractive approach, because unlike the unlimited number of words there are only specific POS categories. If a model can be estimated for each POS tag, then it will be possible to apply this model across different sentences and across different emotions.

From the analysis of the distribution of acoustic features conditioned on individual POS tags for various emotion types it was observed that they do not fit into a particular distribution in any simple way. For that reason, instead of considering the individual emotion-dependent variations in the acoustic features, we decided to concentrate on acoustic feature *differences* between emotions. In contrast, the patterns of the differences were easier to fit into statistical distributions. As can be seen from the example plot (Fig. 2), it was found that a normal distribution can be used to approximate the differences between acoustic features well (as determined by statistical chi-square tests with 5% significance level). Note that although we present the results for only some particular POS categories, the normal distribution approximation was tested on all POS categories and in most cases it provided a good fit to the data. In Table 1, the approximated mean (μ) and standard deviation (σ) values for some tags and emotional pairs are shown as examples. It is important to note that the normal distribution fit to the data is only an approximation. For some pairs it is a very good fit (e.g., Fig. 2a), while for some other cases it provides only a moderate fit (e.g., Fig. 2i).



(a) neu-ang JJ (b) neu-hap JJ (c) neu-sad JJ

(d) neu-ang JJ (e) neu-hap JJ (f) neu-sad JJ

(g) neu-ang JJ (h) neu-hap JJ (i) neu-sad JJ

Fig. 2. The histogram and the approximated normal distribution curve for emotional feature differences for adjectives (JJ). Plots (a), (d), and (g) are for $Neutral - Angry$; (b), (e), (h) for $Neutral - Happy$; and (c), (f) and (i) for $Neutral - Sad$ differences in energy maximum (a,b,c), F0 mean (d,e,f) and tag durations (g,h,i).

			E. Max		F0 Mean		Tag Dur.	
			μ	σ	μ	σ	μ	σ
JJ	n	a	-.093	.278	-37.06	56.41	-.066	.143
JJ	n	h	-.048	.254	-24.95	62.37	-.043	.148
JJ	n	s	-.083	.260	-10.52	58.97	-.071	.211
NN	n	a	-.062	.259	-38.60	62.59	-.059	.144
NN	n	h	-.056	.260	-31.90	68.40	-.034	.142
NN	n	s	-.061	.246	-17.39	64.48	-.045	.170
VB	n	a	-.051	.252	-47.84	66.86	-.029	.131
VB	n	h	-.041	.285	-34.69	72.58	-.019	.133
VB	n	s	-.027	.273	-15.14	70.47	-.024	.142
RB	n	a	-.079	.270	-40.36	57.66	-.051	.162
RB	n	h	-.036	.260	-33.10	62.51	-.035	.170
RB	n	s	-.036	.262	-16.87	61.69	-.026	.186
PRP	n	a	-.036	.295	-43.55	80.09	-.032	.091
PRP	n	h	-.021	.286	-37.23	84.26	-.018	.096
PRP	n	s	-.014	.289	-17.31	84.72	-.014	.010

Table 1. The normal distribution approximation mean and standard deviation values for modeling the differences in the energy maximum, F0 mean and tag duration parameters of neutral(n)-angry(a), neutral-happy(h) and neutral-sad(s) POS pairs.

5. APPLICATION OF THE STATISTICAL POS MODELS TO SYNTHESIS: FIRST EXPERIMENTS

The statistical approximations presented in the previous section can be used for conversion of neutral speech to emotional speech. For a given input neutral speech file, the procedure consists of finding the target values from the corresponding normal distributions and then modifying the input file accordingly.

For ten randomly selected neutral utterances, not included in the analysis set, the above described algorithm was simply applied (without any additional rules or constraints). TD-PSOLA was used for F0 and duration modifications. Energy contour was modified by scaling the speech signal at the appropriate word locations. Smoothing was performed in time domain to ensure a continuous waveform and smooth transitions at word boundaries. Examining the results, although there were some successful conversions in random cases, in general it was observed that the acoustic feature values of different words were not in harmony (due to the probabilistic nature of parameter value generation), thus masking the perception of target emotions.

With the purpose of reducing the mismatch, the findings discussed in Section 3 and presented in Fig. 1 were used to impose specific restrictions on the target parameter values, which were still generated by the corresponding normal distributions. For example, for neutral to angry conversion, it was required that energy maximum of adjectives (JJ) were higher than neutral adjectives. And the target parameter generation from the corresponding normal distribution was utilized until this requirement was satisfied. Having such specific (hard) constraints on the acceptable parameter values improved the results, especially for neutral to sad and neutral to angry conversion. However, still the quality was not completely adequate.

Considering the probabilistic nature of the target feature value generation, such results are not surprising. They indicate that in order to apply any statistical feature generation techniques the differences between neighbor word features should be minimized. Similar to the automatic text generation where higher order N-grams, such as bigram, trigram, or quadrigram models have increasing power, we can expect the presented probabilistic approach to work better for bi-POS or tri-POS models. In addition, speaker dependent models can also improve the conversion performance. Although such models will require significant amount of data, they will have smaller variances and will provide better harmony between target parameters, thus better conversion. These are the topics that we plan to address in the future research.

6. CONCLUSIONS

A statistical approach for modeling the differences in the POS level acoustic features of matching happy, angry, sad and neutral utterances was presented. It was shown that Gaussian probability density functions can be used to model the differences in the statistics of energy, F0 and duration parameters at the POS level for various emotions.

Application of these probability density functions in conversion of emotional speech showed that modeling of a single POS tag is not sufficient and that additional models that take neighboring words into account (similar to N-gram models in speech recognition) should be considered.

Pairwise comparisons of POS level features revealed that, in most cases, the normalized mean and median energy of sad POS tags are larger than neutral, happy or angry POS tags, indicating important differences in energy distribution in emotional utterances. In addition, it was observed that POS based analysis can be helpful for differentiating between happy and angry, and neutral and sad emotions, which often have resembling acoustic features. In summary, the study shows that explicitly accounting for linguistic constraints is essential within any statistical modeling approach to emotion synthesis.

7. REFERENCES

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

[2] S. Lee, E. Bresch, J. Adams, A. Kazemzadeh, and S. Narayanan, "A study of emotional speech articulation using a fast magnetic resonance imaging technique," in *InterSpeech ICSLP*, Pittsburg, PA, 2006.

[3] M. Bulut, C. Busso, S. Yildirim, A. Kazemzadeh, C. M. Lee, S. Lee, and S. Narayanan, "Investigating the role of phoneme-level modifications in emotional speech resynthesis," in *Proc. of Eurospeech, Interspeech*, Lisbon, Portugal, 2005.

[4] T. Dutoit, *An Introduction to Text-to-Speech Synthesis*. Kluwer Academic Publishers, 1996.

[5] J. E. Cahn, "The generation of affect in synthesized speech," *Journal of the American Voice I/O Society*, vol. 8, pp. 1–19, July 1990.

[6] R. Tsuzuki, H. Zen, K. Tokuda, T. Kitamura, M. Bulut, and S. Narayanan, "Constructing emotional speech synthesizers with limited speech database," in *Proc. of ICSLP*, Jeju, Korea, 2004.

[7] G. Kochanski, E. Grabe, J. Coleman, and B. Rosner, "Loudness predicts prominence: Fundamental frequency lends little," *JASA*, vol. 118(2), Aug. 2005.

[8] D. Wang and S. Narayanan, "An acoustic measure for word prominence in spontaneous speech," *IEEE Transactions on Speech and Audio Processing*, 2006.

[9] E. Charniak, "A maximum-entropy-inspired parser," in *Proc. of NAACL*, 2000.

[10] M. P. Marcus, B. Santorini, and M. A. Marcinkiewicz, "Building a large annotated corpus of English: The Penn treebank," *Computational Linguistics*, vol. 19, pp. 313–330, 1993.

[11] L. R. Rabiner and R. W. Schafer, *Digital processing of speech signals*. Prentice-Hall, 1978.