Analysis of emotional speech prosody in terms of part of speech tags

Murtaza Bulut, Sungbok Lee[1]*, Shrikanth Narayanan*[1]

Department of Electrical Engineering, [1]Also, Department of Linguistics
University of Southern California, Los Angeles, CA, USA
mbulut@usc.edu, sungbokl@usc.edu, shri@sipi.usc.edu

Abstract

Representation of emotions in terms of acoustic features of well defined lexical elements is desired for development of emotional speech processing systems. For that purpose, in this paper, the interaction between emotions and part of speech (POS) tags is investigated. Utterances from 3 speakers in angry, happy, sad, and neutral emotions are used to statistically analyze the effects of emotion, POS tag type, position of the tag, and speaker factors on tag duration, energy, and F0 variables. It is found that the main effects of emotion, tag type, and position are significant. Results also show that the effect of emotion is significantly dependent on position, but not on POS tag type. The effect of position is noticeable. POS tags located in the first half of sentences have shorter durations, higher energy, and higher F0 values.

Index Terms: emotional speech, part of speech (POS), prosody

1. Introduction

Research on perception of emotions shows that different words have different emotional effect on human listeners [1]. Therefore one may expect that while expressing different emotions people select and (acoustically) modulate different words differently. The interplay between the linguistic role of words, and their expressive modulation in spoken language is complex, and not completely understood. In this paper we concentrate on speech signal property differences, specifically on prosody parameters (duration, F0, and energy) as a function of part of speech (POS) tags (lexical categories where words in each class grammatically behave the same) in emotion. The specific goal is to investigate the interaction between emotions and POS tags and the factors influencing this interaction.

POS tags are attractive for many reasons. First of all, they are few in number ($\sim$36) and can be identified with high accuracy (more than 90%) with no manual intervention. They are generated and used during text to speech (TTS) synthesis for various purposes such as phrase break assignment [2], prosody generation, homographic disambiguation and target cost calculations [3]. Also, they provide information about word prominence [4]. In addition they are popularly used in natural language processing (NLP) for semantic disambiguation and summarization. If there is a relation between emotion styles and POS tags, and if this relation can be parametrized, it can be used to introduce emotion information in spoken language processing.

In emotional speech research, POS information has been used in emotional speech recognition [5] and only partially (for content/function word discrimination) in emotional speech synthesis [7]. In an effort to investigate their usage for emotional speech resynthesis, in [6] we compared the probabilities of observing differences in the POS tag parameter values across emotions. In this paper, we investigate the effect of these differences in terms of emotion style, part of speech tag type, speaker, and position in a sentence. This is done using a 4-factor (4x13x2x3) mixed-design ANOVA (with repeated measures on emotion style) analyzing duration, energy maximum and median, and F0 median and range dependent variables. As stated above, the purpose is to gain insights regarding, if and, how POS information can be useful in emotional speech processing.

2. Experiment design

In this section we first describe the emotional dataset we use and then give details of how statistical tests were performed.

2.1. Emotional dataset

The results in this study are based on 72 sentences uttered by 3 speakers (who were native speakers of English) in four emotion styles (happy, angry, sad, neutral). The sentences were designed to be semantically neutral, however the distribution of different POS tags was not specifically controlled. The same sentences were used for the all three speakers. Of these, speaker 1 was a male speaker in his late twenties with no professional acting experience. Speakers 2 and 3 were females in their late twenties with degrees from a theater school.

The speakers were asked to utter each sentence in 4 emotion styles, which were happiness, anger, sadness and neutral (i.e., no particular intended emotion) styles. The recordings were performed emotion by emotion, that is all sentences were first recorded in one style then in another. Listening tests were conducted with naive listeners (minimum 4 listeners per file) to evaluate the expressed emotions. The tests were designed as forced choice tests where listeners had to choose one of happy, angry, sad, neutral, or other options. Results showed that in more than 80% of instances, the perceived emotion matched the intended emotion.

The average sentence length was 6.81 words. The 0.25, 0.5, and 0.75 quantiles for sentence word counts were 5,6, and 8 words, respectively. Word boundaries were estimated automatically using HMM models trained using the HTK software.

2.2. Repeated measures design on emotions

In order to analyze the effect of emotions on prosodic features of part of speech tags, a 4 factor mixed design (4x13x2x3) ANOVA was used. There were four independent variables. On one of these, intended emotion, we had repeated measures. There were four levels of the intended emotion: Happy, angry, sad and neutral. The three other independent variables, which were used as between-subject variables, were POS tag type, position of the tag in sentence, and speaker.

2.2.1. *Independent variables*

The POS tag type we used had 13 levels: Adjectives (JJ, 25 (the number of JJ tags in all 72 sentences)), singular or mass nouns (NN, 70), proper singular nouns (NNP, 17), plural nouns (NNS, 5), personal pronouns (PRP, 59), possessive pronouns (PRP$, 17), adverbs (RB, 37), base form verbs (VB, 17), past tense verbs (VBD, 12), gerund or present particle verbs (VBG, 17), past particle verbs (VBN, 9), non-3rd person singular present verbs (VBP, 10), person singular present verbs (VBZ, 7). Our concentration was mainly on POS tags of content words because they were shown to be more prominent than function words [4]. A similar distinction was utilized during the synthesis of emotional speech in the Affect Editor [7].

The position variable had two levels and it was used to mark whether the word was in the first or second half of a sentence. For example, for a 5 (or 6) word sentence, the first 3 words were considered as belonging to first half and last 2 (3) words to the second half. Position was included as a factor as a result of our analysis prior to the design, where it was found that position dependent differences in the acoustic parameter values were significant.

The speaker variable had 3 levels.

2.2.2. *Dependent variables*

The dependent variables were (POS) tag duration, tag energy maximum, tag energy median, tag F0 median, and tag F0 range. Speaker dependent normalizations were performed on all variables in order to minimize effects of speakers. For each speaker the mean and standard deviation of parameter values were calculated at the utterance level across all emotions. These values were used to normalize the tag level parameters. In order to ensure fair competition among the tag energies, the energy values were also normalized at the utterance level so that each utterance had energy median equal to 1.

The normalizations were performed as follows: (1) tag duration = tag duration/mean (utterance duration/utterance word number), (2) tag energy maximum = (tag energy maximum / utterance energy median) / mean (utterance energy maximum / utterance energy median), (3) tag energy median = tag energy median / utterance energy median, (4) tag F0 median = (tag F0 median - mean(utterance F0 median))/std(utterance F0 median), (5) tag F0 range = tag F0 range / mean (utterance F0 range).

The F0 contour was calculated using Praat software. As an energy measure, the average amplitude energy was used. In this method, instead of squares (as in RMS energy calculation) the absolute values are summed over a shifting short time window (Hamming window of 0.015 seconds long). Both F0 and energy contours were smoothed with median filters of length 3.

3. ANOVA analysis results

In this section we present the results of ANOVA tests. The implications of these results are discussed in more detail in section 4.

Results of the ANOVA tests are reported in Table 1 and Table 2. The values are the results of Greenhouse-Geisser statistical test (which was preferred in order to account for the violations of the sphericity assumption) performed using the SPSS software.

In addition to the statistics, figures are also provided to visualize the effect of each parameter. The plots in Fig. 1 are helpful to visualize the main effects of emotion and tag type, and also their interaction.

The main effect of position (and also of emotion) is shown in Fig. 2 where the values of dependent variables are plotted for two conditions: position=1 (i.e., only considering the tags in the first half of sentences) and position=2 (i.e., tags in the second half) conditions. The figure is also helpful to visualize the interaction between emotion and position.

In order to provide a tag level discrimination for emotions, contrast tests were performed to analyze the differences between all 6 emotional pairs. Since position was calculated to be an important factor, these comparisons were performed separately for position=1 and position=2. The results for some POS tags (JJ, NN, VB, RB, PRP) are reported in Table 2. The table also shows the results of post hoc test for POS tag type variable.

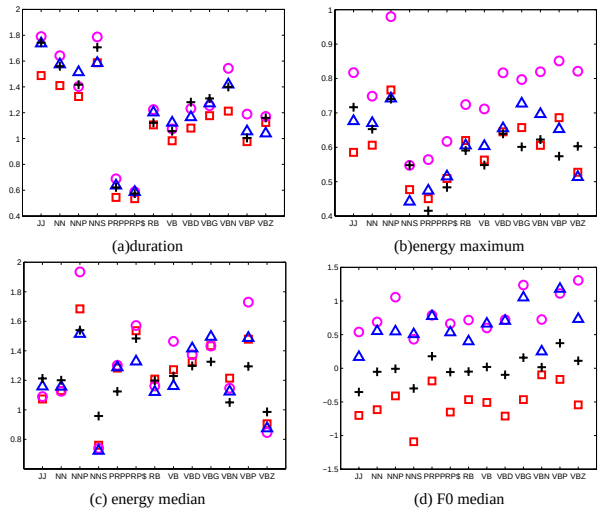

Figure 1: *Emotion and tag type interaction for each dependent variable. Each symbol represents a different emotion: Neutral: □, anger: ◯, happiness: △, sadness: +. The figures show the main effect of POS tag type and emotion, and emotion*tag interaction.*

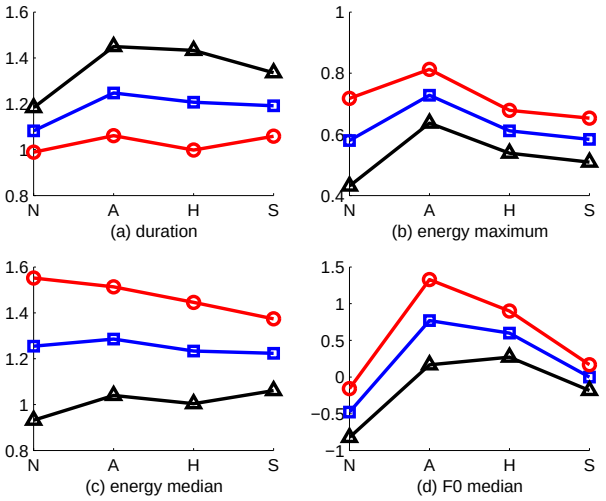

Figure 2: *Shows the main effect of emotion and position. Cases where position=1, position=2, and position=1 or 2 are marked with ◯, △, and □, respectively. Symbols N, A, H, S represent neutral, angry, happy, and sad emotions, respectively.*

Variables	Duration	Energy maximum	Energy Median	F0 median	F0 range
emo	F(2.75,2279.08)=21.53 p<.001	F(2.85,2362.06)=47.47 p<.001	F(2.79,2312.69)=2.60 p=.054	F(2.84,2354.07)=110.00 p<.001	F(2.78, 2301.44)=9.71 p<.001
tag	F(12,828)=46.82 p<.001	F(12,828)=15.01 p<.001	F(12,828)=3.41 p<.001	F(12,828)=1.89 p=.032	F(12,828)=12.11 p<.001
poz	F(1,828)=18.22 p<.001	F(1,828)=67.77 p<.001	F(1,828)=60.04 p<.001	F(1,828)=40.24 p<.001	F(1,828)=5.59 p=.018
spk	F(2,828)=.02 p=.970	F(2,828)=4.32 p=.014	F(2,828)= 0.54 p=.582	F(2,828)= 2.79 p=.062	F(2,828)=5.65 p=.004
emo*tag	F(33.03,2279.08)=1.17 p=.220	F(34.23,2362.06)=1.29 p=.122	F(33.51,2312.69)=1.44 p=.049	F(34.11,2354.07)=.92 p=.596	F(33.35,2301.44)=1.77 p=.004
emo*poz	F(2.75,2279.08)=6.91 p<.001	F(2.85,2362.06)=7.75 p<.001	F(2.79,2312.69)=3.65 p=.014	F(2.84,2354.07)=10.81 p<.001	F(2.78,2301.44)=1.38 p=.25
emo*spk	F(5.50,2279.08)=44.73 p<.001	F(5.70,2362.06)=18.99 p<.001	F(5.58,2312.69)=1.32 p=.249	F(5.68,2354.07)=15.34 p<.001	F(5.56,2301.44)=4.07 p=.001
tag*poz	F(12,828)=2.73 p=.001	F(12,828)=4.17 p<.001	F(12,828)=3.72 p<.001	F(12,828)=3.27 p<.001	F(12,828)=1.67 p=.069
tag*spk	F(24,828)=.29 p=1.00	F(24,828)=.39 p=.996	F(24,828)=.33 p=.999	F(24,828)=.77 p=.776	F(24,828)=.67 p=.866
spk*poz	F(2,828)=2.39 p=.093	F(2,828)=.91 p=.400	F(2,828)=.33 p=.718	F(2,828)=6.79 p=.001	F(2,828)=4.87 p=.008

Table 1: *Greenhouse-Geisser test statistics for a 4-factor mixed-design ANOVA experiment are shown. Independent variables are emotion(emo, 4 levels), tag type (tag, 13), position (poz, 2) and speaker (spk, 3). Significant results are highlighted.*

3.1. Analysis of tag duration, energy and F0 contours

The results in Table 1 show that main effects of emotion (Fig. 1 and 2), tag type (Fig. 1), and position (Fig. 2) were significant on all dependent variables (except the main effect of emotion on energy median, which was insignificant because of the performed utterance level energy normalizations). The main effect of speaker was either not significant or small. This was mainly due to the speaker level normalizations.

Significant effects of emotion and tag type were expected. However, it was particularly interesting to observe the strong effect of position. As it can clearly be seen in Fig. 2, the tags located in the first half of sentences (i.e., position=1) had shorter durations, higher energy, and higher F0 values than the tags in the second half (i.e., position=2). In general, this was true for all tags. As expected, there were significant differences between patterns of differences, due to position, in the values of duration, energy maximum and median, and F0 median, for some tags (i.e., tag*position interaction was significant).

The effect of position was significant for all emotions (Fig. 2). Note also that the effect of position was emotion dependent (i.e., emotion*position interaction was significant) for all variables, except F0 range. This can be seen from Fig. 2 and Table 1.

The results show that we do not have enough evidence to conclude that the effect of emotion on certain parameters – duration, energy maximum, and F0 median – is tag dependent. For energy median and F0 range parameters it was found that the effect of emotions are significantly dependent on tag type. It is an indication that, emotion change affected some tags more than the others. Note, however, that the size of effect was small.

Analysis of 3-way interactions (not shown in the tables), emotion*tag*position, and emotion*tag*speaker, showed that they were insignificant for all acoustic features except duration. For duration, the interaction was small but significant (emotion*tag*position: F(33.03, 2279.08)=1.58, p=0.019, emotion*tag*spk: F(66.06, 2279.08)=1.37, p=0.026).

3.2. Post hoc tests for emotions and POS tags

Contrast analyses of emotions (Table 2) show that emotions can be differentiated from each other at the POS tag level. Moreover, there were differences between information (for emotion differentiation) inherent in different tags. For example, verbs (VB) had less emotion related information than the other tags. These results should be interpreted with caution, however. Considering the relatively small size of the analyzed dataset, they need to be tested for larger databases as well.

The patterns of differences between emotions were dependent on position (Table 1 and Fig. 2). From Table 2 we observe that, in general, for different acoustic features, differences between emotions were more consistent in the second half of sentences. For example, in the second half, a relation (between emotions) observed for duration variable was also observed for energy maximum and F0 median.

As emphasized in section 4, the results discussed in the previous two paragraphs may be specific to the type of dataset (i.e., same sentences for all emotions) that was used in this experiment. It will be interesting to see if these relations still hold for emotional data where all emotions are not expressed with the same sentences.

4. Discussion

Results show that emotions can be differentiated at the POS level. This was an expected result, because many emotional speech analysis studies show that word, syllable and phoneme features change with emotional style change. The more interesting result was to note that emotion and tag type interaction was either insignificant (duration, energy maximum, F0 median) or small (energy median, F0 range). This means that the effect of emotion on duration, energy maximum and F0 median acoustic features was not significantly dependent on POS tag type. However, as shown in Table 2, the emotion related information inherent in different POS tags was different. For instance, verbs (VB) provided less information than adjectives (JJ) or nouns (NN).

In order to understand the interaction between POS tags and emotions better, for randomly selected 10 sentences (i.e., 4x10 utterances), we asked 2 native English speakers to label the most prominent two words. The purpose was to investigate how the position of prominence changed with emotion. The analysis based on position showed that, in general, the same words were salient across different emotions. This means that the prominence marks mostly fell on the same POS tags even when emo-

	JJ	NN	VB	RB	PRP	All tags	Post Hoc test for POS tags
Duration, poz=1	—	h<a/s	h<s	—	n/h<a	n/h<a/s	PRP/PRP$/RB<JJ/NN
poz=2	*n<a/h/s*	*n<a/h/s, s<a*	*n<h*	*n<a/h, s<h*	*n<a/h/s, s<h/a*	*n<a/h/s, s<a*	VB/VBD/VBG/VBP/VBZ<JJ/NN
							PRP/PRP$/VB<NNP/NNS
							RB/VBP/VBZ<NNS
							PRP(PRP$)<all but PRP$(PRP)
Energy maximum, poz=1	n/h<a	s<a	—	h/s<n/a	n/h/s<a, s<n/h	s<n, n/h/s<a	NN/NNS/PRP/PRP$/RB/VB<NNP
poz=2	*n<a/h/s, h/s<a*	*n<a/h/s, h/s<a*	*n<a*	*n<a/h/s, s<a*	*n<a/h, s<a*	*n<a/h/s, h/s<a*	PRP<RB/VB/VBD/VBN/VBP
							PRP/PRP$<JJ/NN/VBG
Energy median, poz=1	—	—	h<a	—	s<n/a/h	h/s<n	VB/VBN/VBZ<NNP
poz=2	*n<a/h/s*	*n<a/s, a<s*	—	*n<s*	*n<h/s*	*n/h<a*	VB/VBN/VBZ<NNP
							JJ/NN/NNS/RB/VBZ<PRP$
							NNS<VBD/VBG/VBP
							VBZ<VBG/VBP
F0 median, poz=1	n<a/h, h/s<a	n<a/h/s, s<h, h/s<a	n<a/h, s<a	n<a/h, s<a/h	n<a/h, s<a/h	n<a/h/s, h/s<a, s<h	JJ<PRP/VBG/VBP
poz=2	*n<a/h/s, s<a/h*	*n<a/h/s, s<a/h*	*n<a/h, s<h*	*n<a/h/s, s<h*	*n<a/h, s<a/h*	*n<a/h/s, s<a/h*	
F0 range, poz=1	—	n<a/h/s	n<a/h	n<a/h/s	—	n<a/h/s	RB/VB<JJ/NN, VBD<JJ
poz=2	*n/s<h*	*n<a/h/s*	*s<a/h*	*n<h/s*	—	*n<s*	PRP/PRP$<JJ/NN/NNP/NNS/VBN
							PRP$<VBZ, VB<NNS/VBN

Table 2: *Contrast analysis of emotions and post hoc comparisons of POS tags. Only significant pairs are shown. Results for position=1, and position=2 are highlighted and in italic, respectively, for easy differentiation. Note that some tags are more helpful than the others to differentiate between emotions. Also note that, the patterns of differences between emotions are, in general, more consistent in the second half.*

tion changed. It is interesting to note that these preliminary analyses were in accord with the results (for the interaction between POS tag and emotion for different acoustic features) in this paper.

Not observing any significant interaction between emotions and POS tags may be due to the type of dataset used. Note that, in this experiment, in order to have well controlled comparisons between emotion, expressions of different emotions were constrained to specific semantically neutral sentences.

Another important conclusion of this study is that position of words (and consequently, of POS tags) was a significant factor. It is seen that the POS tags in the first half of sentences had shorter durations, higher energy maximum and median, and higher F0 median values than POS tags in the second half of sentences. One reason for having lower F0 values toward the end may be because declarative sentences were used. A reason for having lower energy values (in the second half of utterances) may be because speakers uttered a single sentence. Therefore, toward the end of utterance, the subglottal pressure may have decreased and so did energy. Similar results on duration reported in [9] suggest that there may be an "inherent effect of word position in a segment on its duration".

From Tables 1 and 2 we note (as expected) that main effect of tag type was significant for all dependent variables (and especially for duration, energy maximum and F0 range parameters as can be seen from the size of F values). This was because different POS tags carry different prominence information [4] and prominence is related to loudness, duration and F0 [8].

5. Conclusions

In order to better understand how emotions can be parametrized at the utterance level, the relations between 4 emotions and 13 POS tags were analyzed. It was found that the main effects of emotion, POS tag type, and position were significant. Especially, the effect of position was noticeable, indicating that second half of utterances may contain more emotion specific information than the first half. In general, there was not enough evidence to conclude that the effect of emotions was dependent on POS tag type. However, it was observed that emotion related information inherent in different tags was different. The results indicate the complex nature of effects of emotions on POS tag parameters. Analyses of larger emotional corpora with more balanced POS tags presence, and analyses of new type of emotional data, where emotion expressions will not be constrained to same sentences for different emotion, are planned in the future studies to compliment the results presented in this paper.

Acknowledgments: Work supported in part by NIH (DC03172) and the US Army.

6. References

[1] C. Whissel, "The dictionary of affect in language," in *Emotion: Theory, research and experience: Vol.4, The measurement of emotions*, R. Plutchik and H. Kellerman, Eds. New York: Academic, 1989.

[2] A. W. Black and P. A. Taylor, "Assigning phrase breaks from part-of-speech sequences," in *Eurospeech*, Rhodes, Greece, 1997.

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

[4] D. Wang and S. Narayanan, "An acoustic measure for word prominence in spontaneous speech," *IEEE Transactions on Speech and Audio Processing*, vol. 15(2): 690-701, Feb. 2007.

[5] A. Batliner, S. Steidl, B. Schuller, D. Seppi, K. Laskowski, T. Vogt, L. Devillers, L. Vidrascu, N. Amir, L. Kessous, and V. Aharonson, "Combining efforts for improving automatic classification of emotional user states," in *IS-LTC*, Ljubljana, Slovenia, 2006.

[6] M. Bulut, S. Lee, and S. Narayanan, "A statistical approach for modeling prosody features using POS tags for emotional speech synthesis," in *ICASSP*, Honolulu, Hawaii, April 2007.

[7] 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.

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

[9] J. Yuan, M. Liberman, and C. Cieri, "Towards an integrated understanding of speaking rate in conversation," in *Interspeech*, Pittsburgh, Pennsylvania, 2006.