INFLUENCE OF PROSODIC FEATURES AND SEMANTICS ON SECONDARY EMOTION PRODUCTION AND PERCEPTION

Jesin James, Catherine I. Watson, Hywel Stoakes

Department of Electrical, Computer, and Software Engineering, The University of Auckland
jjam194@aucklanduni.ac.nz, c.watson@auckland.ac.nz, h.stoakes@auckland.ac.nz

ABSTRACT

For current speech technology, the emotion expression between machines and humans is made possible by dialog modelling (semantics) and acoustic modelling (prosodic features) of speech. Prosodic features alone are considered sufficient to express and perceive primary emotions. With current focus on social robots, there is also the need to synthesize and recognize nuanced secondary emotions. As the secondary emotions are subtle, this study aims to quantitatively assess (via syllable-level prominence features) whether both semantics and prosodic features contribute to their production and perception. Observations show that the effect of semantics on the prosodic features are significant for the production of the secondary emotions. But unlike primary emotions, there is a need for lexical and grammatical information to support the prosodic component enabling people to perceive the secondary emotions. Additionally effects of English language familiarity have been analysed based on the results of a large scale human perception experiment.

Keywords: secondary and primary emotions, semantics, prosodic features, perception, prominence

1. INTRODUCTION

Emotional speech synthesis and recognition becomes crucial in current technologies that interact with humans as social conversations contain many emotions. Emotions can be classified as primary and secondary emotions. Primary Emotions are innate to support fast and reactive responses. Eg: angry, happy. They are strong reactions to situations and hence they are easy to identify. Secondary emotions arise from higher cognitive processes, based on an ability to evaluate preferences over outcomes and expectations. Eg: relief, hope [1, 2, 12]. They are subtle in nature and hence difficult to identify. Humans express emotions through lexical choices and variations in intonation, speed of delivery, loudness in congruence with the words. For text to speech synthesizers the semantics (lexical choices) is managed by dialog modelling and the other variations are implemented by prosody modelling. To develop emotional speech it is necessary to understand how people use semantic cues and prosodic features to perceive and produce emotions. Table 1 summarizes some studies that explore this. These studies use sentences which are semantically neutral, with participants judging the emotion they can perceive. Some of the studies which explored the effect of linguistic ability on emotion perception [7, 11] have concluded that emotion perception is affected by language knowledge and universal principles of expressions. Also, three of the studies [7, 10, 11] have analysed prosodic features to study their differences across the emotions. Most of these studies have looked at the primary emotions. They have established that the primary emotions (stronger emotions) can be recognized effectively from the prosody component alone even if the speakers spoke pseudo sentences with correctly employed prosodic variations in English [10] and Argentine Spanish [8]. These results are reflected in emotion synthesizers and recognizers that give good accuracy and performance by using prosody variations alone [15, 16]. But compared to the stronger primary emotions, nuanced secondary emotions are more commonly used in social conversations and are needed for human interaction with computer software agents and robots [3]. Hence, speech technology needs to model the secondary emotions as well. This paper uses the contrast in the results of a perception test conducted for primary and secondary emotions to analyze the effect of semantics and prosodic features on the production and perception of secondary emotions.

2. SECONDARY EMOTIONS PERCEPTION

To incorporate a wide array of emotions used by humans in social conversations, the authors of this work developed an open-source emotional speech corpus with 5 primary emotions (angry, happy, neutral, sad, excited) and 5 secondary emotions (anxious, apologetic, enthusiastic, worried, pensive). The corpus (called JLCorpus) contains 2400 sentences spoken by 2 male and 2 female professional New Zealand English (NZE) speakers. The recording of the corpus, perception tests and prosody anal-

Table 1: Previous research works studying the effect of semantics and prosodic features on primary emotions

Experiment & Reference	Emotions	Finding	Features studied
Speakers produced pseudo-utterances for each emotion. Perceived emotions were judged by native listeners. [8]	anger, disgust, fear, neutral, happy, surprise, sad	All emotions were recognized from vocal cues at levels exceeding chance. Anger, sad & fear were most accurate.	Mean F_0 , F_0 Range, Speech rate
Same sentences used for all emotions. Theoretical emotion predictions & acoustics of emotions were studied [10]	anger, sad, joy, fear, disgust	Disgust was poorly recognized, average recognition accuracy for other emotions = 62.8%	Articulation rate, Intensity, F_0, Spectral energy
Argentine Spanish speaker recognized emotions from pseudo-utterances in native language & 3 foreign languages. [7]	joy, sad, anger, fear, disgust	Emotions decoded at accuracy exceeding chance. Emotion perception is language-independent, uses universal principles.	Feature analysis was not conducted.
20 English-speaking listeners judged the emotive intent of utterances. Verbal content was neutral but prosodic elements conveyed 4 emotions. [11]	joy, anger, sad, fear	Identification accuracy was above chance for all emotions. Emotional prosody is decoded by a combination of universal and culture-specific cues.	Mean F_0, F_0 Range, Mean and range of intensity, event density

ysis are explained in [2]. A human perception experiment with 120 participants was conducted to evaluate the corpus. 60 participants each evaluated the primary and secondary emotions separately. Among the participants, 50 were first language NZE speakers (L1) and the remaining 70 were bilingual English speakers (L2). The primary emotions had 1200 sentences which were semantically neutral (eg: "Jack views an art piece"), and the secondary emotions had 1040 semantically neutral sentences and the rest were emotionally coloured (eg: "You should be proud of yourself", for the *enthusiastic*). The participants were given a set of sentences marked with the intended emotion category at the start of the test (training set). This was to familiarise the participants to the various emotions. During the experiment, each participant evaluated 6 sets of 10 sentences each. Each set was shown to the listener on a computer screen. They listened to each of the sentences (audio) in the set and dragged and dropped the sentence icon to the emotion category they perceived the sentence to be. Five emotion category options with a "none of these" option were given to the participants. Once each set was completed the participants proceeded to the next set, until the end of the test was reached. The process was self-timed by the listener, and they could hear the sentences multiple times if needed. They could go back to the training page and hear the training sentences if needed.

Table 2: Perception test results summary.

Emotion Type	Sentence Type	L1/L2	Perception Accuracy
Primary	All	All	69%
Primary	All	L1	70%
Primary	All	L2	68%
Secondary	All	All	40%
Secondary	Emotionally coloured	L1	54%
Secondary	Emotionally coloured	L2	67%
Secondary	Semantically neutral	L1	41%
Secondary	Semantically neutral	L2	33%

Table 2 summarizes the results of the perception test differentiating the emotion type, sentence type and L1/L2 participant categories. The overall perception accuracies for primary and secondary emotions were 69% and 40% respectively. For the semantically neutral primary emotion sentences, both L1 and L2 speakers performed almost equally. This is in alignment with the findings in previous works [7, 11] that the perception of primary emotions is not heavily affected by semantics and language knowledge. Unlike the primary emotions, the table shows that there is contrast in results between emotionally coloured and semantically neutral sentences for the secondary emotions. This contrast in the results between primary and secondary emotions, and also scarcity of studies on secondary emotions has motivated further analysis. The contrast in perception accuracies of L1 and L2 participants was investigated in detail. For emotionally coloured sentences, the L2 have performed better than L1, while for the semantically neutral case, the L1 have performed better than L2. The English familiarity alone was not found to have significant effect on the perception accuracy for secondary emotions [$F(1, 58)=1.04$, $p = 0.3$], while the effect of semantics was found to be significant on the perception accuracy [$F(1, 114)=962.9$, $p = 0$]. This indicates that both L1 and L2 participants had similar difficulty in correctly perceiving the secondary emotions, while the presence of semantic information significantly affected participants' perception accuracy. To analyse if the semantic knowledge and the English familiarity of the participants (L1/L2) has any interaction effects on the perception accuracy of secondary emotions, a 2-factor ANOVA was conducted. The combined effect of L1 vs L2 and semantic influence on the perception was found to be significant [$F(2, 112)=16.26$, $p = 0$]. Also a post-hoc Tukey test was conducted to understand which all categories are

contributing to this significant interaction. From the test it was seen that both L1 and L2 have interaction effects on the semantically neutral and emotionally coloured sentences. Since the interaction effect is significant (p=0), the perception accuracy obtained cannot be generalized for both L1 and L2 participants under varying influence of semantics. For primary emotions even without semantic information the emotion perception accuracy is above chance. But, the effect of semantics on the perception accuracy of secondary emotions observed in this study is different to what has been observed for primary emotions. This result is critical when developing emotional speech synthesizers with secondary emotions. The semantic information has to be developed such that it is congruent with the emotion to enhance the perception accuracy of the these subtle emotions. In the next section we try to understand whether the semantics has an influence on the production of these secondary emotions.

3. SECONDARY EMOTIONS PRODUCTION

In this section we look at the effect of semantics on the production of secondary emotions by the speakers who spoke these sentences for the JLCorpus. The effect on emotion production is analysed by studying the impact of semantics on prosody features like F_0, RMS energy and duration. These features were chosen because during a wider acoustic analysis they were identified as most significant in distinguishing these emotions. Only the results of F_0 are discussed here. Similar trends were observed for the other two features as well. A previous study by the authors [4] looked into sentence level analysis without differentiating the vowels and the consonants. A sentence level analysis results in a large variation of features due to interaction of the consonants. Also, previous studies show that vowels are most likely to contain the emotion information [9]. To reduce the effects of averaging out prosody features, here we analyze the vowels separately.

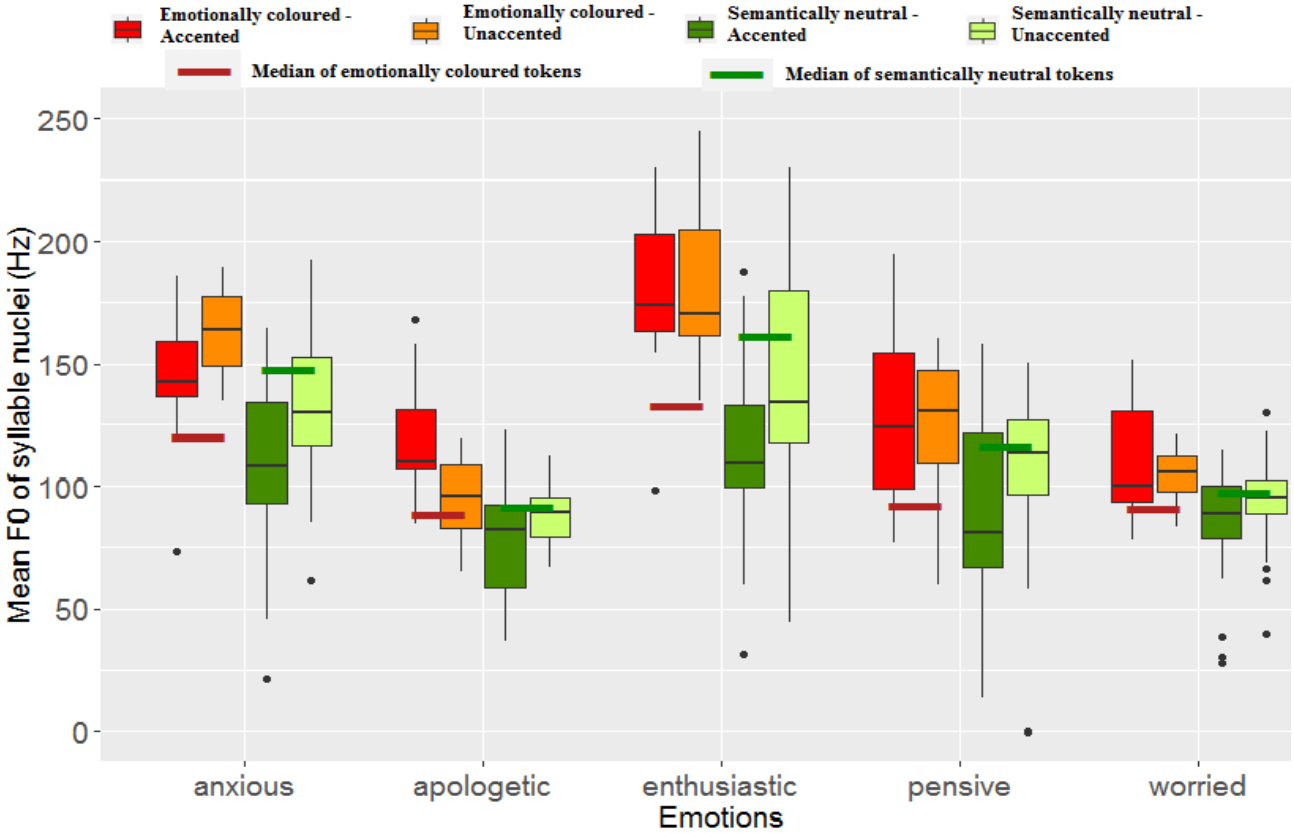
Mean F_0 for the vowels were calculated separating the emotionally-coloured and semantically neutral sentences for 4 speakers individually. Considering all the vowels may neutralise the effect of key prosodic features like heightened prominence on particular syllables. Prominence is the property by which linguistic units are perceived as standing out from the sentence [13]. The accented syllable is taken as the indicator of prominence here. A subset (360 sentences - 60 for each emotion of male2 speaker with a total of 821 syllable tokens) of the corpus was segmented at the phonetic level using WebMAUS [5], and the boundaries were hand-corrected where required. The accented syllables and their nuclei were marked via perceptual decision making using Praat as the visualising tool. Mean F_0, RMS energy and duration of the accented syllable nuclei were taken as the measure of prominence as used in [13, 14]. The effect of emotions on mean F_0 of the accented and unaccented syllable nuclei were analyzed separately. Both accented and unaccented syllables have statistically significant effects of semantics based on the results obtained from ANOVA (for accented syllables - *[F(1, 240)=18.51, p = 0]* and for unaccented syllables - *[F(1, 558)=29.87, p = 0]*). The boxplot in Figure 1 shows the distribution of the mean F_0 values of emotionally coloured and semantically neutral accented and unaccented tokens. The shades of red are emotionally coloured tokens (first 2 of every emotion set), and shades of green are semantically neutral tokens (last 2 of every emotion set). The first and third boxes (darker color) for each emotion are the accented syllable tokens and the other 2 (lighter colour) are unaccented tokens. The dark red and dark green lines indicates the overall median F_0 value for the emotionally coloured tokens and semantically neutral tokens respectively for each emotion. The lines show that the emotionally coloured tokens have lower median value of the feature mean F_0 compared to the semantically neutral tokens for all emotions (for all emotions dark green line is higher than dark red line). Instead of generalising this for all tokens, we separated them into accented and unaccented tokens. It is visually clear that the accented tokens have a larger emotion separation, and this aligns with past studies that prominent syllables are good emotion indicators. We can see that the accented tokens have similar mean F_0 regions (dark red and dark green boxes have larger overlap) for each emotion. While the unaccented tokens show larger difference between the emotionally coloured and semantically neutral tokens (light red and light green boxes have less overlap). Hence, it is clear that it is the difference in the mean F_0 of the unaccented tokens that is contributing to the difference in the overall mean values of the emotionally coloured and semantically neutral tokens (dark red and dark green lines).

A post-hoc Tukey test (Table 3) revealed that only *apologetic* emotion was effected by the semantic influence in the accented syllables. Hence the effect of semantic influence is not strong for the accented syllables. Now for the unaccented tokens, the emotion separation is not as good as the accented tokens. But the difference between emotionally coloured and semantically neutral is visually evident. Also, a Tukey test (Table 3) confirms this as 4/5 emotions analysed

Table 3: Results for semantics and emotions on Mean F_0 as feature (**Bold** indicates statistical significance)

Syllable type	Groups being compared	Diff. in mean	Lower Conf.	Upper Conf.	p-adj
Accented syllable nuclei	sem.neutral:anxious-coloured:anxious	2.88	-0.53	6.29	0.19
	sem.neutral:apologetic-coloured:apologetic	**-4.04**	**-7.75**	**-0.32**	**0.02**
	sem.neutral:enthusiastic-coloured:enthusiastic	-0.67	-4.311	2.97	0.99
	sem.neutral:pensive-coloured:pensive	0.28	-3.36	3.92	1
	sem.neutral:worried-coloured:worried	0.77	-2.77	4.31	0.99
Unaccented syllable nuclei	**sem.neutral:anxious-coloured:anxious**	**3.06**	**0.54**	**5.57**	**0.00**
	sem.neutral:apologetic-coloured:apologetic	2.49	-0.35	5.35	0.15
	sem.neutral:enthusiastic-coloured:enthusiastic	**4.93**	**1.82**	**8.04**	**0.00**
	sem.neutral:pensive-coloured:pensive	**7.04**	**4.47**	**9.60**	**0.00**
	sem.neutral:worried-coloured:worried	**4.01**	**1.22**	**6.70**	**0.00**

Figure 1: Accented and Unaccented syllable nuclei Mean F_0 vs semantics



had statistically significant results. This implies there is a difference between the accented and unaccented syllables that is affected by the emotional colouring of the sentence. The presence of emotionally coloured words allows the speaker to impart prominence to those words specifically (containing accented syllables), while lowering the prominence feature (mean F_0 here) of the other syllables. For semantically neutral cases there are no semantic cues for words on which the prominence has to be imparted. Thus there is not much differentiation between the accented and unaccented syllables.

4. CONCLUSION

Past studies and JLCorpus perception test show that the primary emotions perception without semantic information is better than chance. Secondary emotions are essential for Human Computer Interaction applications. An in-depth analysis of these emotions is conducted here, and effect of semantics and prosodic features on the production and perception of secondary emotions was studied. It was seen that the speakers use the semantic cues and impart differences among accented and unaccented syllables during emotion production. This will be an important consideration while developing secondary emotions corpora. Also, the language familiarity of the participants and the semantics affect the perception accuracy. The results will be useful for dialog modelling and prosody modelling of secondary emotions for emotional speech synthesis.

5. REFERENCES

[1] Damasio, A., 1994, Descartes' error, emotion reason and the human brain, *Grosset/Putnam, pp 135-139*

[2] Jesin James, Li Tian, Catherine Inez Watson, 2018, An Open Source Emotional Speech Corpus for Human Robot Interaction Applications, *in Proc. Interspeech*

[3] Jesin James, Catherine Inez Watson, Bruce MacDonald,2018, Artificial Empathy in Social Robots: An analysis of Emotions in Speech, *In Proc. IEEE International Conference on Robot and Human Interactive Communication*

[4] Jesin James, Catherine Inez Watson, 2018, The role of prosody and semantics in the perception of Secondary emotions, *in Proc. ProsLang Workshop on the Processing of Prosody across Languages and Varieties*

[5] Kisler, T.,Schiel, F., Sloetjes, H., 2012, Signal processing via web services: the use case WebMAUS, *in Proc. Digital Humanities, pp 30-34*

[6] Kitayama, Shinobu and Ishii, Keiko, 2002, Word and voice: Spontaneous attention to emotional utterances in two languages, *COGNITION AND EMOTION, 16(1):29â59*

[7] Pell, M.D., Monetta, L., Paulmann, S. et al., 2009, Recognizing Emotions in a Foreign Language, *in J Nonverbal Behav, Volume 33, Issue 2, pp 107â120*

[8] Marc D. Pell, Silke Paulmann, Chinar Dara, Areej Alasseri, Sonja A. Kotz, 2009, Factors in the recognition of vocally expressed emotions: A comparison of four languages, *Journal of Phonetics, Volume 37, Issue 4, pp 417-435*

[9] Pereira C, Watson C.I., 1998, Some Acoustic Characteristics of Emotion, *In Proc. ICSLP*

[10] Scherer, K.R., Banse, R., Wallbott, H.G. et al.,1991, Vocal cues in emotion encoding and decoding, *Motivation and Emotion, Volume 15, Issue 2, pp 123â148*

[11] Thompson, W. Balkwill, L., 2006, Decoding speech prosody in five languages, *Semiotica, 407-424*

[12] Christian Becker-Asano, Ipke Wachsmuth, 2010, Affective computing with primary and secondary emotions in a virtual human, *Autonomous Agents and Multi-Agent Systems, vol 20, pp 32*

[13] Jacques Terken, 1991, Fundamental frequency and perceived prominence of accented syllables, *in The Journal of the Acoustical Society of America 89(4 Pt 1):1768-76*

[14] Agaath M.C. Slugter, Vincent J. van Heuven, 1996, Acoustic correlates of linguistic stress and accent in Dutch and American English, *in International Conference on Spoken Language Processing*

[15] D.-n. Jiang, W. Zhang, L.-q. Shen, L.-h. Cai, 2005, Prosody analysis and modeling for emotional speech synthesis, *in Proc. IEEE Int. Conf. Audio Speech Signal Process. (ICASSP), pp. 281-284.*

[16] C.H. Wu, W.B. Liang, 2011, Emotion recognition of affective speech based on multiple classifiers using acoustic-prosodic information and semantic labels, *IEEE Transactions on Affective Computing, vol. 2, no. 1, pp. 10-21 .*