EXPRESSIVE SPEECH SYNTHESIS: SYNTHESISING AMBIGUITY

Matthew P. Aylett[1 2], *Blaise Potard*[2], *Christopher J. Pidcock*[2]

University of Edinburgh, Informatics[1]
CereProc Ltd.[2]
Edinburgh, UK
matthewa@inf.ed.ac.uk

ABSTRACT

Previous work in HCI has shown that ambiguity, normally avoided in interaction design, can contribute to a user's engagement by increasing interest and uncertainty. In this work, we create and evaluate synthetic utterances where there is a conflict between text content, and the emotion in the voice. We show that: 1) text content measurably alters the negative/positive perception of a spoken utterance, 2) changes in voice quality also produce this effect, 3) when the voice quality and text content are conflicting the result is a synthesised ambiguous utterance. Results were analysed using an evaluation/activation space. Whereas the effect of text content was restricted to the negative/positive dimension (valence), voice quality also had a significant effect on how active or passive the utterance was perceived (activation).

Index Terms: speech synthesis, unit selection, expressive speech synthesis, emotion, prosody.

1. INTRODUCTION AND STATEMENT OF RELATION WITH PRIOR WORK

In many systems, speech synthesis is required purely to communicate neutral dynamic information to a user, for example their bank balance or the time of an appointment. However, as computer applications become more complex, for example by simulating environments, or taking on the role of a trainer or tutor, the interaction required with users also becomes more complex. In such systems, user engagement becomes more important, and in order to build systems which can create a compelling sense of engagement, Human Computer Interaction (HCI) research has begun to look at alternatives to the dominant approach of user-centered design. Two alternatives to the traditional HCI approach are ludic design, which focuses on the importance of encouraging playfulness in a design[1], and experience-centred design, which focuses on the sense of experience that a system would like to engender in a user[2]. In these design approaches, ambiguity, normally avoided in interface design, can be harnessed to encourage intrigue, mystery and delight[3]. Speech synthesis is a key enabling technology for pervasive design, and in order to face the new challenges of affective, eyes-free and mobile systems, speech synthesis technology needs to offer designers the flexibility and functionality that can support these new design methodologies. This presents a challenge for speech synthesis, both in terms of creating ambiguity in synthetic utterances, and in evaluating this ambiguity.

Ambiguity is often the result of a tension between opposing perceptions. It is this tension which can add to a user's curiosity and engagement. This is quite different from neutrality, where there are no dominant or contrasting perceptions. For example 'hot and cold' is ambiguous, whereas 'warm' is neutral. In natural speech, ambiguity is often used to create a specific effect, for example irony. One definition of irony is *an expression or utterance marked by a deliberate contrast between apparent and intended meaning.[4]* One method used by human speakers to generate irony, is to use a contrasting emotion to the content spoken, for example "What a brilliant day" said with an angry or stressed voice. Contrasting meaning and emotion in this way creates a complex picture of the speaker. It conveys more than the straightforward utterance "What a horrible day", because the tension between the voice and the content suggest a complex internal state which in turn adds to the sense of character.

Current speech synthesis systems typically produce neutral speech, although more recently, work in expressive speech synthesis has examined how to create speech which *unambiguously* conveys an emotion or an underlying expressive goal. This work has a long tradition of focusing on evaluating a distinct set of between three and nine, extreme, sometimes termed *primitive* emotional states, such as disgust, fear, anger, joy, sadness, and surprise [5]. This presents a problem for creating ambiguous utterances, because a very strong emotion in the voice will dominate the perception of the utterance. Instead a more controlled approach is required which can offset other features in the utterance. The CereVoice speech synthesis system uses a distinct set of sub-corpora containing different *voice qualities* to achieve a more subtle change in the perceived emotion in an utterance.

Voice quality is an important factor in the perception of emotion in speech[6]. However, unlike speech rate and pitch, which can be modified relatively easily using digital signal processing techniques such as PSOLA, modifying voice quality is more difficult, especially if it is important to retain naturalness. Rather than modifying speech to create the effect, an alternative approach is to record different voice qualities and use them directly during concatenative synthesis. This approach has been applied to diphone synthesis [7] and has been extended to unit selection in the CereVoice system which uses pre-recorded voice quality sub-corpora in unit selection [8]. This is different from other unit selection approaches which have instead examined the use of sub-corpora of specific emotions, e.g. [9] where Happy, Angry and Neutral sub-corpora were incorporated into an emotional voice in Festival. By focusing on voice quality rather than specific emotions, CereVoice allows a combination of DSP techniques and unit selection to craft a more varied and subtle set of speech styles[10].

As with [7] three styles of voice quality (VQ) are available: Neutral (the default for the recorded corpora) and two sub-corpora of lax (calm) and stressed (tense) voice quality. Adding an XML tag in the speech biases the selection of the units to come from the sub-corpora. However, the extent to which this unit-selection VQ approach suc-

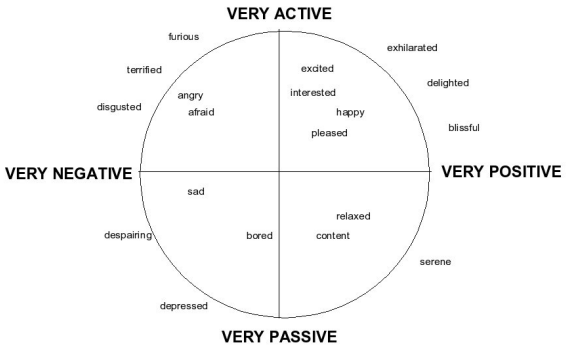

Fig. 1. *Activation/Evaluation Space*

ceeds in conveying a negative/positive perception of the utterance has not been formally evaluated until now.

In order to both evaluate this approach, as well as evaluate the success or failure of creating an ambiguous utterance, we require an evaluation methodology which allows a response to be shifted depending on competing factors. In this work, we adapt the approach taken by FEELTRACE[11] and evaluate utterances within the *activation/evaluation space*.

FEELTRACE was developed specifically for assessing gradual changes in emotion by allowing subjects to place the emotion in a two dimensional space called the evaluation/activation space. This space is based on previous work in psychology [12, 13] and regards emotions as having two components, a valence which varies from negative to positive, and an activation which varies from passive to active (See Figure 1). In this way, rather than asking subjects which emotion they perceive in an utterance, the subject chooses a point in this two dimensional space. This approach is especially powerful for detecting shifts in emotion.

2. RESEARCH QUESTIONS

In order to create a conflict between voice quality and text content, sentences were chosen with content intended to be both negative and positive. Neutral sentences, and natural speech with neutral, lax and stressed voice qualities, were used as controls.

Our research questions were:

RQ1: Does voice quality change equate to a change in the positive/negative (valence) perception of an utterance?

RQ2: If so, can we use a mismatch between voice quality and text content to create ambiguity in synthetic utterances?

3. METHODOLOGY

Voice quality is one feature among many that effect the perception of emotion in speech. As we wished to discover the effect of *voice quality change* only, we used the same approach as Hofer et al [9], and asked subjects to rate the emotion in the synthetic and natural speech by choosing a position in the activation/evaluation space (Figure 1). We also asked subjects to rate naturalness on a 5 point scale (Bad/Poor/Fair/Good/Excellent). The experiment was carried out online (see Figure 2) using 14 English native speakers. Subjects were requested to use headphones. There were three factors in the experiment: Synthesis/natural speech **SYN**, Stressed/Lax/Neutral voice quality **VQ**,

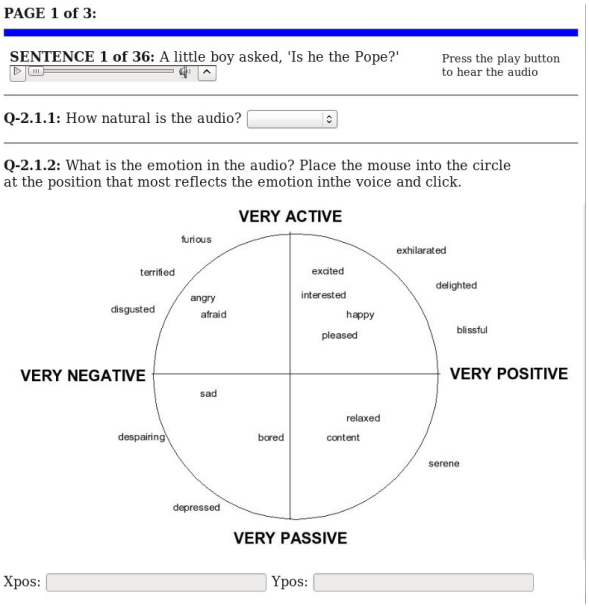

Fig. 2. Online Experimental Setup

and Positive/Negative/Neutral Text content **TCONT**. Although sentences were present for all conditions in synthetic speech, natural speech sentences were missing for: Positive Text with Stressed Voice Quality, Negative Text with Lax Voice Quality and Positive/Negative Text with Neutral Voice Quality. Synthesis was generated using the CereProc Sarah RP female voice with natural stimuli held out from the speech database.

Positive and Negative text content was selected from online news materials by evaluating sentences using the dictionary of affect[14]. The dictionary of affect gives valence and activation scores to 8742 emotional words. Positive sentences were selected which contained more positive words and the converse for negative sentences. All sentences were manually checked in order to ensure the overall semantic meaning also matched the desired text category. There were 12 sentences in each category.

Our hypotheses were:

1. H1: There is a significant difference in perceived valence between utterances with stressed and lax voice qualities, for both natural and synthetic speech.
2. H2: Text content has a significant influence on perceived valence.
3. H3: Where text content mismatches voice quality, the perceived valence moves towards the neutral point in the valence scale due to the opposing perceptions creating by the ambiguity.

4. RESULTS

4.1. Naturalness

It is important to assess naturalness when testing speech synthesis in any context as very poor naturalness will confound other perception results and call into question the utility of any process that reduces naturalness below an acceptable level.

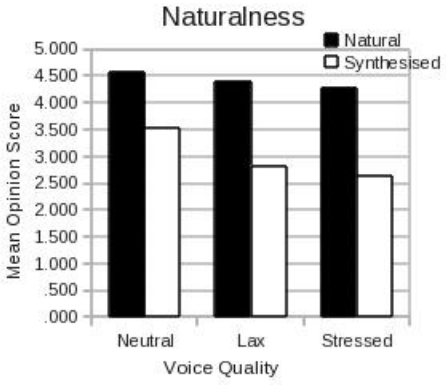

Fig. 3. Mean opinion score by voice quality and by natural and synthetic speech.

Figure 3 shows naturalness expressed as a mean opinion score (MOS) grouped by voice quality. A grouped univariate ANOVA analysis with two factors, **SYN** and **VQ** was carried out. Both factors were significant (SYN: $F(1, 462)=201.98$, $p<0.001$), (VQ: $F(2, 462)=11.00$, $p<0.001$) with a just significant between factor interaction ($F(2, 462)=3.13$, $p<0.05$). Post-hoc Tukey tests showed a significant drop in naturalness for stressed and lax synthesised sentences compared to Neutral sentences suggesting that concatenating mixed sub-corpora causes an increase in synthetic artifacts. However this drop in quality is typically less than 0.5 MOS. The naturalness of neutral utterances compares favourably with state of the art systems (typically around 3.5)[15]. Results using MOS scores should be treated with care as there is a strong argument that the underlying subject data should not be treated as parametric data. However MOS is a default standard in speech synthesis and using MOS allows a multifactor analysis of the data using a grouped ANOVA analysis. Although MOS data is rarely Gaussian, an ANOVA analysis is acceptable based on the sampling theorem providing each cell has sufficient data points (commonly 10 or above).

4.2. Voice Quality

Figure 4a shows the mean values for neutral sentences with different voice quality within the activation/evaluation space by **SYN** and **VQ**. The means of natural utterances are shown in black, synthetic utterances in grey.

A grouped multivariate ANOVA analysis was significant for voice quality, for both valence ($F(2, 186)=16.75$, $p<0.001$) and activation ($F(2, 186)=57.48$, $p<0.001$), with a significant interaction between **VQ** and **SYN** for activation ($F(2,186)=6.03$, $p<0.005$). Post-hoc tests showed that, in general, natural sentences with lax and stressed voice quality were rated further from the centre of the activation/evaluation space than synthesised sentences. However synthesised sentences showed similar, if less marked, effects of voice quality than natural sentences.

Although one aim of voice quality change is to modify the valence, there is also a strong effect on the perception of activation. Lax voice quality is associated with low activation and positive valence, and stressed voice quality is associated with high activation and negative valence.

In addition, we must note that our neutral voice quality was rated very positively. For commercial systems, voice talents are chosen for having pleasant positive voices and this can undermine the use of a neutral voice quality in such voices as a representative control.

However, results show that altering voice quality affects the perception of valence and activation in an utterance. Although the effect for valence is significant only between stressed and other voice qualities this allows to accept hypothesis H1.

4.3. Effect of Text Content

Due to missing cells for text content in the natural stimuli set, a second ANOVA was carried out on synthetic materials only. A grouped multivariate analysis was carried out with **VQ** and **TCONT** factors. **VQ** results supported those for the neutral sentences ($F(2,325)=18.15$, $p<0.001$ for valence and $F(2,325)=53.25$, $p<0.001$ for activation). **TCONT** had no significant effect on naturalness or activation but showed a significant effect for valence ($F(2,157)=10.43$, $p<0.001$). Post-hoc Tukey tests showed a significant difference between negative text content and neutral and positive content ($p<0.05$) but not between neutral and positive text content. Means for text content for neutral voice quality utterances only are shown in Figure 4b.

Results show that text content affects the perception of valence in a synthesised utterance. This allows us to accept hypothesis H2.

4.4. Effect of Irony

Figure 4c shows the means for matching (black), mismatching (dark grey) and neutral utterances (light grey). There is a clear shift of the mismatching utterances towards the neutral area in the evaluation space. This shift was significant (Tukey post-hoc test $p<0.05$) for Lax voice quality.

Looking more closely at the distributions of these five categories (See Figure 5), we can see a marked difference between matching and mismatching/neutral utterances.

Subjects responses are constrained to be within (or almost within) the activation/evaluation circle. This puts a limits on possible differences in variance by constraining the tails of the distributions. This results in skew for distributions with off centre means, hence the different distribution shapes of non-ambiguous (matching) utterances, from ambiguous (mismatching) utterances and neutral utterances.

Overall we see a significant shift towards neutral valence for the Lax/-ve utterances and a non-significant tendency for Str/+ve utterances to also move towards neutral valence. We have already shown that both voice quality and sentence content affect valence, therefore we can conclude that we have succeeded in creating ambiguous utterances where contrasting features are creating an element of tension allowing us to accept hypothesis H3.

5. CONCLUSION

We have shown that the use of the activation/evaluation space is a useful and effective means of evaluating valence shifts caused by competing features in spoken utterances. We have also shown the impact of voice quality on perceived emotion in terms of valence and activation and how the text content of the utterance also modifies the perception of valence.

Furthermore, a combination of voice quality associated with positive valence and text content associated with negative valence creates a mismatch which produces a perceived valence closer to the neutral part of the scale. As we have significant evidence that these

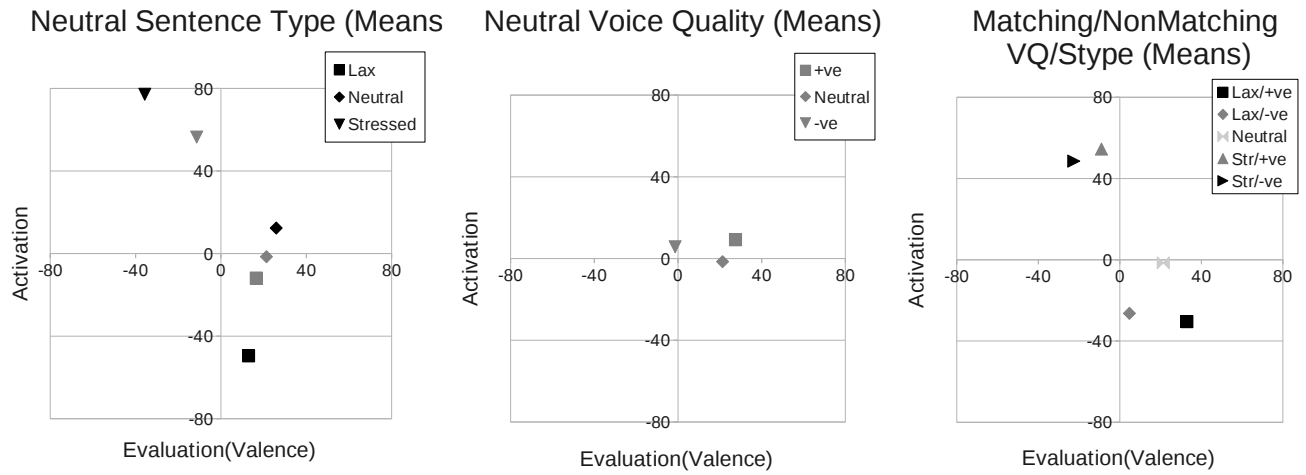

Fig. 4. a) Mean values for neutral sentences by different voice qualities. Black - Natural Speech, Grey - Synthetic Speech. b) Mean values by text content for neutral synthesised utterances. Positive sentences - '+ve', negative sentences - '-ve'. c) Mean values by matching/nonambiguous (black), mismatching/ambiguous (dark grey) and neutral (light grey) synthesised utterances.

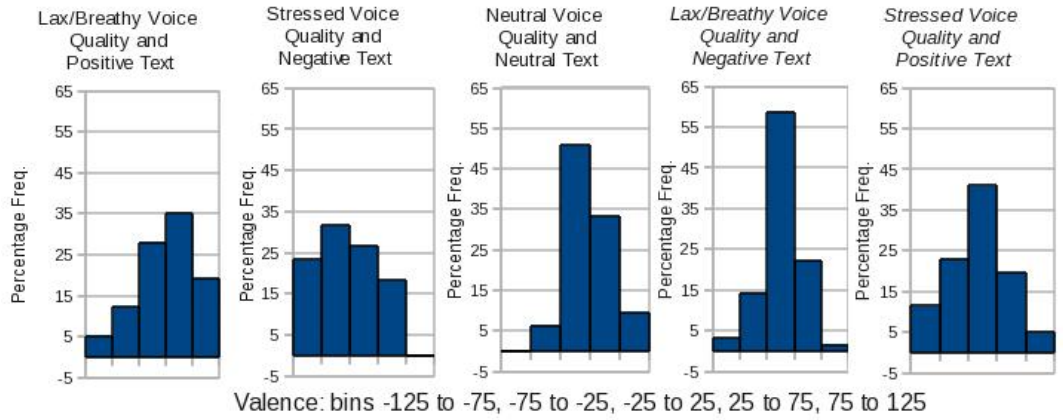

Fig. 5. Comparison of the distribution of valence results. **TCONT/VQ** mismatching/ambiguous utterances in italics, (left). matching/non-ambiguous conditions (right) and responses to neutral condition (centre).

features produce an effect on valence in isolation, together they are creating a tension in the utterance and producing an emotionally ambiguous stimuli.

However, a limitation of our activation/evaluation space approach is that it can't distinguish between a neutral utterance and an ambiguous one. Subjects were asked to give a single response and Figure 5 shows that conflicting features do not create a bi-modal response but are instead merged.

However, qualitatively the mismatched utterances do not sound like the neutral utterances. We have made an example of all nine conditions available on the internet [1] and encourage the reader to listen to the differences.

Although this is strong indirect evidence of creating ambiguous utterances, we would like to have a more explicit way of testing the difference between the neutral and the ambiguous. This requires more advanced evaluation methodologies which can deal with issues such as motivation, intention and conversational function. We believe the results presented here offer a good starting point for investigating these higher level responses to synthetic speech stimuli. Future work will investigate the direct affect of ambiguity on the perception of character, and, through the assessment of more complete systems, the utility of this approach in increasing engagement.

[1] http://homepages.inf.ed.ac.uk/matthewa/ssw2013VQ/

6. ACKNOWLEDGEMENTS

This research was funded by the Royal Society through a Royal Society Industrial Fellowship.

7. REFERENCES

[1] A. J. Morrison, P. Mitchell, and M. Brereton, "The lens of ludic engagement: evaluating participation in interactive art installations," in *Proceedings of the 15th international conference on Multimedia*, ser. MULTIMEDIA '07, 2007, pp. 509–512.

[2] J. McCarthy and P. Wright, *Technology as Experience*. MIT Press, 2004.

[3] W. W. Gaver, J. Beaver, and S. Benford, "Ambiguity as a resource for design," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ser. CHI '03, 2003, pp. 233–240.

[4] American Heritage, *The American Heritage Dictionary of the English Language.*, 4th ed. Houghton Mifflin Company, 2009.

[5] M. Scrhöder, "Emotional speech synthesis: A review." in *Proceedings Eurospeech 01*, 2001, pp. 561–4.

[6] C. Gobl and A. N. Chasaide, "The role of voice quality in communicating emotion, mood and attitude," *Speech Communication*, vol. 40, no. 1-2, pp. 189–212, Apr. 2003.

[7] M. Scrhöder and M. Grice, "Expressing vocal effort in concatenative synthesis." in *ICPhS*, 2003, pp. 2589–92.

[8] M. Aylett and C. Pidcock, "Adding and controlling emotion in synthesised speech," UK Patent GB2 447 263A, September 10, 2008.

[9] G. Hofer, K. Richmond, and R. Clark, "Informed blending of databases for emotional speech synthesis," in *Proc. Interspeech*, 2005.

[10] M. P. Aylett and C. J. Pidcock, "The cerevoice characterful speech synthesiser sdk," in *AISB*, 2007, pp. 174–8.

[11] R. Cowie, E. Douglas-Cowie, S. Savvidou, E. McMahon, M. M. Sawey, and M. Scrhöder, "Feeltrace': An instrument for recording perceived emotion in real time." in *ISCA Workshop on Speech and Emotion*, 2000, pp. 19–24.

[12] H. Schlosberg, "A scale for judgement of facial expressions." *Journal of Experimental Psychology*, vol. 29, pp. 497–510, 1954.

[13] R. Plutchik, *The Psychology and Biology of Emotion.* New York: Harper Collinns, 1994.

[14] C. M. Whissell, "The dictionary of affect and language." in *Emotion: Theory, Research, and Experience.*, R. Plutchik and H. Kellerman, Eds. Academic Press, 1989, vol. 4, pp. 113–131.

[15] S. King and V. Karaiskos, "The blizzard challenge 2010," in *Blizzard Challenge Workshop*, 2010.