Informed Blending of Databases for Emotional Speech Synthesis

Gregor O. Hofer, Korin Richmond, Robert A.J. Clark

Centre for Speech Technology Research
University of Edinburgh, UK
g.hofer@sms.ed.ac.uk

Abstract

The goal of this project was to build a unit selection voice that could portray emotions with varying intensities. A suitable definition of an emotion was developed along with a descriptive framework that supported the work carried out. A single speaker was recorded portraying happy and angry speaking styles. Additionally a neutral database was also recorded. A target cost function was implemented that chose units according to emotion mark-up in the database. The Dictionary of Affect supported the emotional target cost function by providing an emotion rating for words in the target utterance. If a word was particularly 'emotional', units from that emotion were favoured. In addition intensity could be varied which resulted in a bias to select a greater number emotional units. A perceptual evaluation was carried out and subjects were able to recognise reliably emotions with varying amounts of emotional units present in the target utterance.

1. Introduction

The overall goal of the speech synthesis research community is to create natural sounding synthetic speech. To increase naturalness, researchers have recently identified synthesising emotional speech as a major research focus for the speech community [1]. One way synthesised speech benefits from emotions is by delivering certain content in the right emotion (e.g. good news are delivered in a happy voice), thereby making the speech and the content more believable.

Previous emotional speech synthesisers have focused mostly on modifying parameters of model based synthesisers to make utterances sound emotional. This type of work suffers from the limitations of the synthesis quality of the model based systems. The work carried out during this project is based on unit selection, which in general cannot modify acoustic parameters. Also most work in emotional speech synthesis has aimed to create full-blown emotions or expressions, with the notable exception of Marc Schröder [2], who did not use unit selection because of its limitations in modifying acoustic parameters but used MBROLA instead. The problem with modifying acoustic parameters is that the vocal correlates of emotion are not very well understood [3]. Therefore, unit selection can not only yield better results in naturalness ratings but also in the emotion dimension.

With unit selection one can model a few emotional states as closely as possible. Recent work has used this approach with three full blown emotions [4]. A separate database was recorded for happy, angry, and sad tone of voice. Although this type of method almost guarantees good results, it is impossible to generalise the voice to other types of emotional expression, because the emotion is not explicitly modelled but just reproduced from the database. This also makes it very costly to implement different emotion intensities.

Furthermore, implementing a unit selection voice that can portray varying degrees of emotion is not trivial given the limitations of unit selection. The two options are to record separate databases for each emotion in each intensity or to create a *blending* technique that is able to vary the amount of emotional units selected for a synthesised utterance. This technique should of course be theoretically motivated, meaning the type and number of units should correspond to what is expected by the listener. Blending is a term for an interpolation between separate databases [5]. It allows for more gradual changes between the different styles of the databases.

One of the major criticisms of unit selection is that it can only model speech that is in the database. In this paper we discuss a method for using unit selection to generate emotional speech that uses an underlying emotion model to select the correct units. The technique will be able to vary intensity of a given emotion by selecting units from different databases and joining them.

2. The Emotional Speech Synthesiser

The Festival Speech Synthesis System was used as a basis for the research. It was slightly modified to fit the requirements for an emotional speech synthesiser. It had to take into account the external information of the Dictionary of Affect and an emotional target cost function was added to the already existing target cost. This function used a primitive emotion model to find the right units in the blended databases.

2.1. Emotion Model

The Dictionary of Affect [6] was originally created by Cynthia Whissell and has been used in previous studies [7] for mapping speech to emotion categories. In our research the dictionary was employed as a source for the emotional connotation of a given word. The Dictionary itself, which contains 8742 words, represents emotions in a two dimensional circumplex model with the emotion spaced around a circle. Figure 1 shows the Circumplex model. The two dimensions are called valence and arousal. Valence corresponds to a negative/positive rating and arousal to a mild/intense rating. Each word in the dictionary has a score for these dimensions.

The markup for the synthesised utterances does not use dimensions, it just uses categorical emotion labels, like happy and angry. Therefore a translation between the two schemes had to be developed. As can be seen from Figure 1, categorical happiness and categorical anger have similar arousal values but a "happy" word would have a higher valence than an "angry" word. If the current emotion was happy, words with higher va-

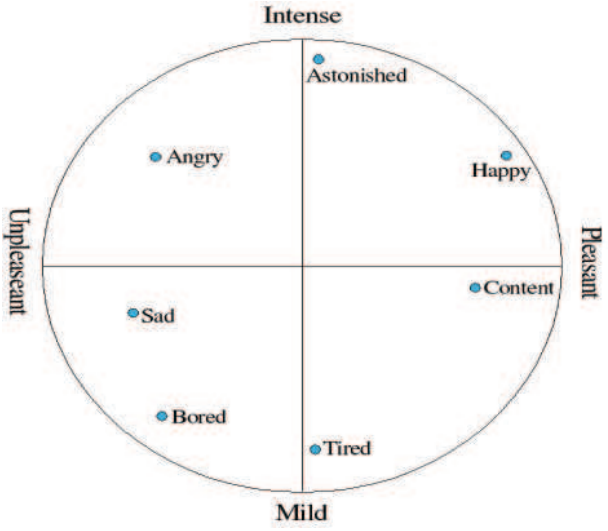

Figure 1: The circumplex model of affect [8]. The horizontal dimension is called valence and the vertical dimension is called arousal. Categorical emotion labels have been added for better orientation. Categorical happiness and categorical anger have a similar arousal score.

lence had a high chance of being synthesised as happy. In a similar way the arousal dimension was considered where arousal meant higher intensity. The higher the arousal of a certain word the more important it is that that word is synthesised with the correct emotion.

2.2. Emotional Unit Selection

The challenge, at synthesis time, was to find the units in the database that best matched the target utterance. Two measures were used to find the best units. The join cost gives an estimate of how well two units join together and the target cost describes how well a candidate unit matches the target unit. [9] gives a good overview of this procedure. Instead of having a separate voice for each emotion, a target cost function was added to the target cost to find the most suitable units for synthesising speech with the desired emotion. The text to be synthesised was also marked up using XML tags. The tags produced an emotion feature in the target utterance. This feature was present at all processing stages. Instead of just comparing the emotion feature of the target utterance to the markup of the candidates, an outside source of knowledge was applied to find the most suitable units. It was hypothesised that only certain words are really important to be synthesised from an emotional database where as others can be synthesised from the neutral database. This lead to another hypothesis where it was possible to vary intensity of a given emotion by varying the amount of words that were synthesised with emotional units.

The additional target cost function was designed to influence which words received emotional units depending on the emotion of the target utterance and the score of the word in the Dictionary of Affect. Each word in the target utterance that was present in the dictionary was mapped onto to the circumplex model by its score. The scores determined what preference was given to that word to be synthesised from emotional units. For example, the word "wife" has a score of 2.7143 in the valence dimension and 1.8333 in the arousal dimension. Each

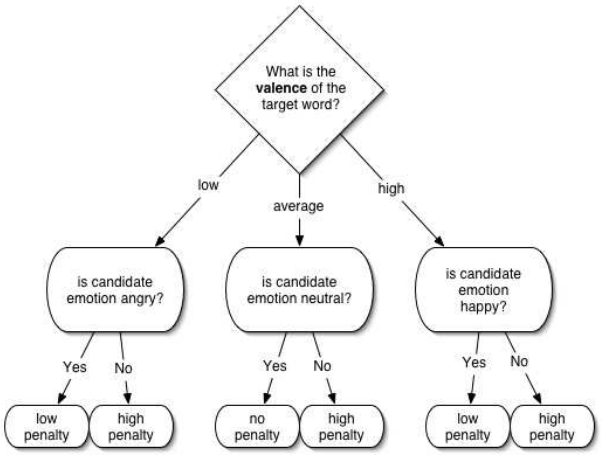

Figure 2: Part of the flow chart of the emotional target cost function.

dimension lies within range 1 and 3. This means that "wife" is a positive word with average intensity. It will have a good chance of being synthesised using "happy" units when the value of the emotion feature is happy. Words with higher arousal would have a higher chance of being synthesised using emotional units. "wearing" on the other hand has a valence score of 2.0000 and a arousal score of 1.8000 which means that this word will most likely be synthesised using neutral units.

Figure 2 shows part of the flowchart of the emotional target cost function. In addition, the algorithm was biased towards selecting units from the same emotion within a word to keep joins low. In general the algorithm was biased towards selecting neutral units. If a word was not in the dictionary or its score was in the opposite direction to the value of the emotion feature present in the target, then the target cost formulation favoured the selection of neutral units. To allow for the synthesis of an emotion with different levels of intensity, penalties were parameterised with weights, these effectively allowed control over the overall proportion of units used from each database. The overall number of words that were synthesised in a certain emotion was varied using different penalty parameter weights. The penalties formed part of the target cost score and increasing the penalties incurred a greater target cost score for using the wrong type of units for emotionally marked words. The trade off between the emotional component of the target cost and the other components and the join cost, resulted in more or less emotional units being chosen depending on the value of the penalty. High penalty values resulted in the use of fewer emotional units in the synthesis of words with average valence scores, and a high valence score was needed before the system would choose emotional units. Lower penalty values resulted in more emotional units being used, even for example, when the arousal score was low.

2.3. Emotional Voices

Three sets of penalties were implemented which resulted in 7 voices that used varying amounts of emotional units. The voices are listed in Table 1. The full_angry, neutral and, full_happy voices used the same sets of penalties. The half_angry and half_happy voices used the same set of penalties and the some_angry and some_happy voices used the same set of penalties. The main difference between the voices was the percentage

emotion	joins	neutral units	ha. units	an. units
full angry	18	0	0	50
quite angry	21	23	2	25
some angry	22	34	1	15
neutral	21	50	0	0
some happy	19	25	25	0
quite happy	19	21	29	0
full happy	19	0	50	0

Table 1: The number of joins and units in each emotion condition for the target utterance: "His wife Zoe was wearing Prada and Gucci while studying Zoology at University"

of units from different databases that were included in a synthesised utterance. With the first set of penalties utterances consisted usually only of units from a single database. Of course in this case the quality of the synthesis was the highest because there were very few joins between units from different databases. Furthermore, the emotions sounded the strongest with this set of penalties.

The next set of penalties was aimed at keeping the balance of units about 50/50 between the neutral database and an emotion database. If the target word was in the Dictionary of Affect and had an emotional connotation that corresponded somewhat to the target emotion, chances were high that it would be synthesised from emotional units. With this set of penalties the synthesis quality was generally worse than with the previous set but there were a few instances when the synthesis quality was improved. Sometimes, units that were missing in the neutral database were present in the emotion databases and this improved the quality of the synthesis. As intended, the emotions did not sound as strong with this set which was intended.

The final set of penalties was designed to use more neutral units by using the minimum amount of emotional units possible. Emotional units were only used if the target word had very emotional attributes in the Dictionary of Affect. The synthesis quality was very similar to the previous set, but it sometimes was very difficult to perceive an emotion. The emotion effect was very weak with this set of penalties. Table 1 shows the distribution of units for an example synthesised utterance over each emotion condition. It was attempted to modulate the intensity of emotions by varying the amount of emotional units used in an utterance. A formal evaluation was conducted to asses the accuracy of the synthesised emotional speech.

3. Data collection

It is not straightforward to obtain emotional speech suitable for use in a unit selection voice. Normally, the speaker is told to speak with a constant tone of voice throughout the recording. During emotional speech, acoustic parameters vary, therefore it is desirable to have control over the speaker to keep the variation low. Two categories of emotional speech in research settings have been identified [3]. One uses natural speech that has been recorded from a speaker during an emotion eliciting event like playing a computer game. The other category uses portrayed emotion by a trained actor. The former method has the advantage that the data is very natural but it is also less controlled. The actor approach has the advantage of allowing control over the emotions portrayed as well as the intensity of the emotion but it has the disadvantage that the data is not completely natural. Actors were used in previous studies on the perception of

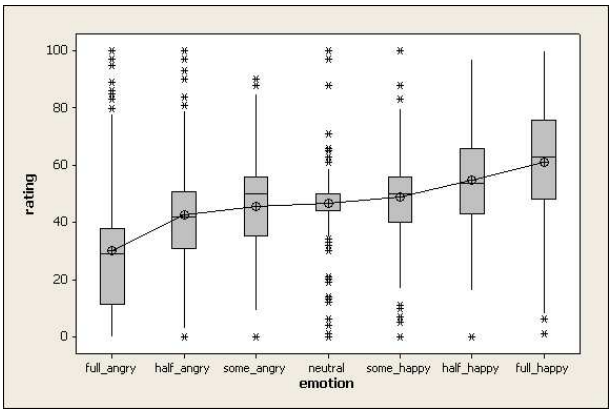

Figure 3: The means and variance of the ratings according to the emotion. Outliers are marked with *

emotional speech [3].

For this research, an actress was recorded portraying 3 different emotions: happy, angry, and neutral. It was concluded that the benefit of having control over the emotions and the benefit of a trained speaker outweighs the disadvantage of not recording completely natural speech. The actress was professionally trained and is regularly employed for dubbing TV shows. We recorded 400 sentences in each emotion. The script consisted of newspaper sentences, which had been selected to optimise diphone coverage. Since we used a trained speaker there was no apparent interaction between the emotion condition and the emotional connotation of the sentences in the script. The three recored speech databases where marked up with an emotion feature and pooled together to form one large database that resulted in a copy of each sentence in all three emotions.

4. Evaluation

The accuracy of the developed emotional synthesis was assessed in a perceptual test. Four different carrier sentences were synthesised in each of the 7 voices. A neutral version of each sentence was also included for control. There was an effort to keep the number of joins constant for one sentence, because too much variation in the number of joins could introduce a *lurking variable* in the evaluation. The total number of synthesised sentences was 28. The rating was performed by the subjects using a continuous scale represented by a slider bar that was labelled angry on the left side, happy on the right side and neutral in the centre. The program recorded the input on a scale from 0 to 100 where 50 meant neutral, 0 meant angry, and 100 meant happy. A continuous response was needed because the strength of a given emotion was part of the assessment. Each subject had to listen to three blocks of 28 sentences over headphones. Each block was randomised in. No way was provided to repeat a sentence. The instructions were given on paper and there was no time limit on the experiment.

The experiment had 13 participants. A graph of the descriptive statistics is shown in Figure 3. Both the mean and the variance are shown for each emotion condition. The mean ratings of the angry emotion conditions (full_angry, half_angry, some_angry) become higher as less and less angry units are present in the stimuli. This rating means that the perceived anger is decreasing with less angry units. The neutral condition had the smallest variance and a mean of 46.6 which is not

emotion	lower	center	upper	
full angry	-20.58	-16.33	-12.08	*
half angry	-8.05	-3.80	0.45	
some angry	-5.40	-1.15	3.10	
some happy	-2.06	2.19	6.44	
half happy	3.94	8.19	12.44	*
full happy	10.17	14.42	18.67	*

Table 2: Results of Fischer's LSD for neutral vs. emotions. Significant differences are marked with *

significantly different from the anticipated mean of 50. The rating means of the happy conditions (full_happy, half_happy, some_happy) become higher as a greater number of happy units are present in the target utterance. Which indicates that more happy units make the perceived utterance seem happier. In general, it can be concluded that the number of emotional units present in the target utterance has an effect on the perceived intensity of the emotion.

Finally, a one-way Analysis of Variance (ANOVA) was carried out to measure the difference between the ratings for each condition. A Fisher's Least Significant Difference (LSD) comparison was carried out to measure individual differences between the mean ratings of the emotions. The results are summarised in Table 2. There were only significant differences between the neutral condition and the full emotion conditions and the half_happy emotion condition. The half_angry condition was only marginally not significant. The little number of significant differences may be attributed to the small sample size.

5. Discussion

Apart from fulfilling the goals set at the beginning, several other interesting possibilities for extensions have opened up during the course of this project. First, the database collection process could be modified to support the speaker in the elicitation of emotions. All databases collected during this project had very similar coverage. It would be much more efficient to have just full coverage for the neutral database. For the emotion databases, a new script would have to be designed in accordance with the Dictionary of Affect to ensure only the units are recorded that are needed to cover a certain emotion. The distribution of units for the utterances synthesised during this project can give some idea on what needs to be recorded for a given emotion. The emotion mark-up in the database could be enhanced by introducing more fine grained labels, or by switching to a dimensional representation. By switching to a finer grained representation of emotions, units from the database can be better matched with the intended intensity of the utterance.

The synthesiser so far can only portray two different emotions, happy and angry. We are aware that there are many more emotions and various subcategories within an emotion like hot and cold anger. These should be considered in future work although it is not clear if this is necessary. It might be enough to have a synthesiser that can portray positive and negative affect in various degrees. The ability to portray a wide range of emotions might not be as important as the ability to portray them in various intensities. For many applications it is the idea of "how good" and "how bad" that might be the most important to communicate.

6. Conclusions

We have described a version of Festival that is able to synthesise emotional utterances in varying intensities. The hypothesis was confirmed that emotions are perceived as more intense when a greater number emotional units are included in the utterance. It has been shown that by varying the amount of emotional units, the intensity of a perceived emotion can be varied. Random selection of emotional units will be carried out in the future to test the effect of the dictionary of affect. There was a small interaction between the sentence and the emotion. Future work will address the creation of better carrier sentences to test emotional speech synthesisers. The general quality of the emotions synthesised was very good, since all tested emotions showed an effect in the right direction with the right intensity. Also, several participants of the rating study mentioned that they were surprised how good the quality of the emotional synthesis was.

It has been argued before that unit selection is a dissatisfactory method from a research perspective because it is not true model of speech but rather plays back previously recorded speech in an intelligent way. Systems that try to model the speech apparatus might be more interesting for research purposes, but as long as there is no clear understanding of the acoustic parameters involved in emotional speech, we believe unit selection synthesis remains the best option for synthesising emotional speech.

7. References

[1] Bailly, G., Campbell, N., Möbius, B. (2003) ISCA special session: hot topics in speech synthesis, Eurospeech 2003.

[2] Schröder, M. (2003). Speech and Emotion Research. An overview of Research Frameworks and a Dimensional Approach to Emotional Speech Synthesis. PhD thesis. Universität des Saarlandes. Saarbrücken.

[3] Banse, R., & Scherer, K. (1996). Acoustic Profiles in Vocal Expression. Journal of Personality and Social Psychology. 70, 614-636.

[4] IIda, A., Campbell, N. Higuchi, F., & Yasumura, M. (2003). A Corpus based speech synthesis system with emotion. Speech Communication, 40, 161-187.

[5] Black, A. W. (2003). Unit Selection and Emotional Speech, Eurospeech 2003.

[6] Whissell, C. M. (1989). The dictionary of affect and language. In Plutchik, R. and Kellerman, H., editors, Emotion: Theory, Research, and Experience. volume 4: The measurement of emotions, pages 113-131. Academic Press, New-York.

[7] Piwek, P., Krenn, B. Schröder, M. Grice, M., Baumann, & S. Pirker, H. "RRL: A Rich Representation Language for the Description of Agent Behaviour in NECA". Proc. of the AAMAS workshop on "Embodied conversational agents - let's specify and evaluate them!", Bologna, Italy. 2002.

[8] Russell, J. A. (1980). A circumplex model of affect. Journal of Personality and Social Psychology. 39, 1161-1178.

[9] Clark, R. A. J., Richmond, K., & King, S. (2004) Festival 2 - Build Your Own General Purpose Unit Selection Speech Synthesiser. 5th ISCA Speech Synthesis Workshop, Pittsburgh, PA.