Transformation of LF Parameters for Speech Synthesis of Emotion: Regression Trees

Michelle Tooher, Irena Yanushevskaya and Christer Gobl

School of Linguistic, Speech and Communication Studies
Trinity College Dublin
{mtooher; yanushei; cegobl}@tcd.ie

Abstract

This paper outlines an approach to modelling the dynamics of voice source parameters as observed in the analysis of emotional portrayals, by a male speaker of Hiberno-English. The emotions portrayed were *happy*, *angry*, *sad*, *bored*, and *surprised*, as well as *neutral*. The voice source parameters extracted from emotionally coloured repetitions of a short utterance – by means of inverse filtering followed by source model matching – were modelled using classification and regression trees. Regression trees were built using the voice source parameters of the *neutral* repetition of the same short utterance, in order to transform the voice source parameters from *neutral* to one of the five emotions. Re-synthesis of emotion-portraying utterances using transformed voice source parameter dynamics resulted in synthesised utterances which were confirmed by listening tests to represent the targeted emotion categories. The results suggest that the addition of dynamic voice source information in parametric synthesis of emotion will improve the quality of emotion synthesis.

1. Introduction

Following from a detailed voice source analysis of a small database of male speech [1], in which the speaker portrayed the basic emotions: *anger, joy, boredom, sadness*, and *surprise*, a number of synthesis implementations and perceptual tests were carried out. The aim was (i) to attempt to verify the results of the analysis through synthesis, (ii) to explore whether changing the source parameter settings while maintaining the neutral filter settings would generate emotionally coloured output, and (iii) to propose a method for modelling the dynamics of voice source parameters in an utterance in a way that can be used to generate synthetic utterances which have emotional colouring. This method aims to model, for the source parameters, the relationships between the *neutral* and the emotion-portraying utterances. The information used for the modelling consists of four source parameters based on the analysis of the above mentioned database (see [1]) as well as information on the relative prominence of the syllable (stressed/unstressed).

Acoustic correlates of emotional speech are often listed in terms of features such as utterance intensity, f_0 contour, and voice quality as well as timing and speech rate. It is well established that voice quality correlates with emotion [2, 3], but the ways in which voice quality is defined tends to vary. The description of voice quality often includes f_0 and perturbation measures such as jitter and shimmer, or labels based primarily on perceptual evaluation such as breathy, creaky, harsh, etc. In synthesis, these qualities can be generated by using a small set of parameters, see, e.g., [3]. In [4], the phonation type (modal, falsetto, breathy, creaky, tense voice) was changed according to the emotion required, along with other factors such as speech rate, mean pitch, and pitch range. Perception tests in [4] showed that synthesised emotions do depend on phonation types, as well as other factors.

Many studies have shown vocal correlates of emotion [2, 3], and suggest emotion related settings of parameters: for example, anger would be characterised as having high mean pitch, narrow pitch range, and tense voice [4]. Carlson et al. [5] found through analysis-by-synthesis experiments that emotions are "signalled through a complex interaction of segmental and prosodic cues", among them the use of durations, f_0, and pitch movement. When synthesising emotional speech with a parametric synthesiser, most frequently global (static) parameter settings are used, ignoring the very complex utterance internal dynamic variation.

Audibert et al. [6] reported on a series of experiments involving monosyllabic words, investigating the relative weight of acoustic parameters in emotion expression. The approach they adopted is somewhat similar to that used in our second experiment presented here. Certain parameters of a neutral utterance were replaced with those of an emotion-portraying utterance using copy synthesis. The parameters analysed included f_0, intensity, phonemic duration, source, residue and vocal tract filter. Different combinations were tested in re-synthesis experiments, ranging from full copy synthesis to synthesis using the source, residue and vocal tract filter of the expressive stimulus, with the phonemic durations, f_0 and intensity extracted from the neutral expressions. They also tested various other combinations of parameters extracted from expressive and neutral monosyllabic utterances.

Our experiments are similar in essence, but rather than using the relatively static settings of short words, we are focussing on the extensive dynamic variation that one finds in a longer utterance. The voice source measures were based on an earlier inverse filtering analysis and source modelling [1]. Three different kinds of output were synthesised using these data, and each was perceptually tested. It is important to note that in this analysis-by-synthesis approach, we allow the voice source parameters to vary as they do in true emotive or neutral sentences, and we do not attempt to generalize or average the dynamics of the voice source.

2. Method

The initial data used for analysis purposes consisted of repetitions of an all-voiced sentence, 'We were aWAY a YEAR ago', produced by a male Hiberno-English speaker in his mid twenties. The speaker, who is not a professional actor, portrayed a number of emotions including *happy, angry, sad, bored*, and *surprised.* A neutral rendition of the utterance was also recorded. Recordings took place in a semi-anechoic chamber and the informant maintained a constant distance of 30 cm from the microphone. Listening tests (outlined in [1]) confirmed the credibility of the targeted emotional portrayals.

2.1. Background analysis

The source analysis was two-tiered. The first tier involved the measurement of voice source and vocal tract parameters. This was carried out by performing semi-automatic closed-phase covariance inverse filtering followed by voice source model matching. The analysis was executed in three phases. Firstly, manual closed-phase marking was performed on the speech waveform. Then, automatic covariance inverse filtering was carried out on the closed-phases, which was subsequently hand-corrected using interactive software for optimising the accuracy of the vocal tract parameters (formant frequencies and bandwidths). Thirdly, the LF (Liljencrants-Fant) voice source model [7] was manually fitted to the inverse filtered waveform. From the source modelling, LF parameter data were obtained. (Note that by "LF parameters" we here mean the standard set of parameter often used with the LF model namely EE (amplitude of the main excitation), RA (effective duration of the return phase normalised to the glottal cycle), RK (glottal pulse symmetry), RG (normalised glottal frequency), and f_0. These are general source parameters and are of course not in any way exclusive to the LF model.)

The second analysis tier involved statistical analysis of the LF parameters for each emotion portrayal. Full details of the analysis performed are outlined in [1]. The results of the statistical analysis suggested that emotion differentiation could be represented in terms of LF parameters, as a unique combination of LF parameter settings was observed for each emotion. It was also concluded that the dynamics of parameters would need to be considered in future analysis and synthesis. That is, it would not be adequate to set static values for each parameter and each emotion in synthesis: the parameters need to be allowed to vary throughout an utterance.

To verify and test the conclusions reached in [1], two synthesis experiments were carried out, as outlined below. The third experiment extended the analysis insofar as it attempts to model the dynamics of the LF parameters through an utterance, assuming one has a neutral baseline and specific emotional targets.

2.2. Synthesis experiments

Three synthesis experiments were undertaken: (i) Copy synthesis, (ii) Hybrid copy synthesis, and (iii) Synthesis from source parameter transformation. All synthesis experiments used the LF source implementation in the KLSYN88a parametric synthesizer [8]. Note that the LF parameters are not directly used in the re-synthesis. Instead, the equivalent parameters in KLSYN88a are used, i.e. AV (Amplitude of Voicing), TL (Spectral Tilt), SQ (Speed Quotient), OQ (Open Quotient) and F0, which can be derived from LF parameters.

Copy synthesis: As a first step in the re-synthesis procedure, simple copy synthesis of emotionally-coloured utterances was performed. The formant frequencies and bandwidths obtained from the inverse filtering and the LF parameter data as derived from the analysis in [1] were provided as input to the KLSYN88a synthesiser for each emotion. The inverse filtering residual was not included.

Hybrid copy synthesis: What we term "Hybrid copy synthesis" was performed to validate the hypothesis, suggested in [1], that emotion-specific combination of LF parameter settings could be useful for the differentiation of emotion and in synthesis of emotion. In hybrid copy synthesis, the formant tracks obtained by inverse filtering the *neutral* utterance of 'We were aWAY a YEAR ago' and the LF parameters of each of the emotions *happy*, *angry*, *sad*, *bored*, and *surprised* were input to the synthesiser yielding five synthesised utterances, each representing one of the above emotions. Because the goal of this experiment was to verify whether specific combinations of voice source parameter settings do encode emotion, in synthesis we wanted to only change the voice source information and maintain the vocal tract filter settings to those of *neutral*. Furthermore, although it is known that there are timing correlates to emotional expression in speech, for the purposes of this experiment we eliminated this variable and retained the timing characteristics of the *neutral* utterance. Consequently, the durations of all emotion-portraying utterances were scaled to that of *neutral*. This was achieved by setting anchor points and performing linear interpolation (as in [1]). The resulting utterances represented synthesised speech that had a *neutral* vocal tract filter configuration, *neutral* timing characteristics but voice source characteristics of the emotional utterances.

Synthesis from source parameter transformation: This experiment tested the possibility of transforming the source parameter trajectories of a *neutral* utterance to emotion-encoding source parameter trajectories. The source parameters were transformed using regression trees (see Section 2.3), one of which had been built for each source parameter. Each parameter regression tree was given the following data as input: the target emotion, the source parameter values of *neutral*, and the information on syllable prominence (stressed/unstressed). The output of the regression trees were the new source parameter values, based on the information supplied. See Fig. 3 for comparison of the original f_0 parameter input and new f_0 parameter output. These new (transformed) source parameters were given as input to the synthesiser with the vocal tract filter formant information of *neutral*.

2.3. Regression trees

The approach in this paper was to model the voice source dynamics in an emotion-portraying utterance as a function of the *neutral* utterance. That is, given a set of source parameters for a *neutral* utterance, would it be possible to transform these parameters into emotion-portraying parameter tracks?

Regression trees were chosen for the modelling of parameter dynamics due to their ability to model relationships between sets of data. At the training or tree building stage, data are recursively partitioned and regression trees are built from these partitions. At each node in the tree, the equality at the node determines whether to take the left branch or the right branch from that node. For example, in Fig. 1, at the topmost node, if the target emotion is either *bored* or *sad* we take the left branch to node number 2, at which point the *neutral* f_0 value is used to calculate a new f_0 value, by multiplying the *neutral* f_0 value and a boolean value for syllable prominence (0 or 1) by regression coefficients and summing the results to a constant value. (Coefficients at the terminal nodes are not displayed here due to space restrictions.) In this way, if we have a set of *neutral* f_0 values as they occur in an utterance, along with the syllable prominence information and the target emotion, a new set of f_0 values can be determined.

A machine learning regression tree algorithm, GUIDE (Generalised Unbiased Interaction Detection and Estimation) [9] was used for the building regression trees. The algorithm fits piecewise linear regression models to the data at each terminal node (grey nodes in Figs. 1 and 2), allowing parameter transformation.

Regression trees were built in the same way for each source parameter employed in the KLSYN88a synthesiser, i.e. for F0, AV, SQ, OQ and TL. Trees could potentially be built in the same way for any of the LF parameters (EE, RK, RG, and RA, as used in [1]) provided they are available in the synthesiser.

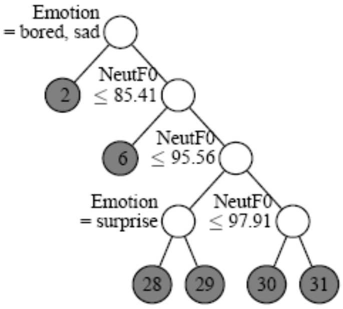

Figure 1: *Regression tree for transforming the source parameter F0 of neutral to that of a target emotion.*

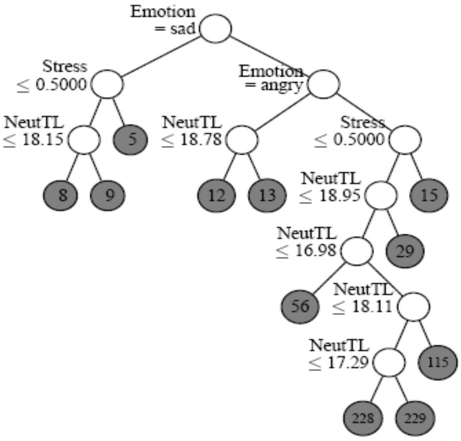

Figure 2: *Regression tree for transforming the source parameter TL of neutral to that of a target emotion.*

Certain knowledge about the source parameters can be acquired by analysing the regression trees built for each parameter. For example, looking at the left branch of the tree in Fig. 1, representing the transformations required for F0 of *bored* or *sad*, it can be seen that the transformation is very basic; there is only one node at which the transformation takes place. However, looking at the tree branches to the right, for *angry*, *happy* and *surprised*, there are more decisions to be made before reaching a terminal node at which the transformation can take place, resulting in a more dynamically varying parameter track. These observations support [5], where it was reported that "the sad utterance was rather monotonous, while the happy and angry showed a more dynamic pitch contour". Similar tendencies were found for other glottal source parameters, see for example Fig. 2.

2.4. Listening tests

Listening tests were conducted for each of the three synthesis experiments to assess the emotional content of the synthesised utterances. Both formal and informal listening tests were undertaken.

Based on informal listening tests performed by the authors, the copy synthesis was found to be of very high quality, with relatively little loss compared to the original recording, producing good replicas of the original portrayed emotions.

Similarly, the hybrid copy synthesis was judged to be of good quality (but obviously the differences compared to the original would be greater in this case), where the intended emotions were clearly conveyed.

With respect to synthesis using the transformed source parameters, informal listening tests conducted by the authors suggested only small differences between the utterances synthesised with *bored* and *sad* as the intended emotion, and between *angry* and *happy* as the intended emotion. It was decided for this reason not to treat *bored* and *sad* as well as *angry* and *happy* separately in the formal listening tests. Instead, these emotions were grouped according to what we term here 'emotion category': *bored/sad* and *happy/angry*, which reflect the activation (or arousal) level that characterises these emotions.

10 participants were asked to listen to the synthesised utterances and to decide which one of the following four emotion categories each utterance could be assigned to: *happy/angry, bored/sad, surprised*, or *no emotion.* 6 synthesised utterances were presented to listeners 5 times in random order (i.e. 30 tokens). The 6 utterances were the 5 emotion-portraying utterances synthesised using the formant parameters of *neutral*, and the source parameters transformed using the regression trees for each emotion. A copy synthesis of the *neutral* utterance was also included in the listening test.

Participants in the formally conducted listening tests assigned *angry* and *happy* to the category *happy/angry* in 87.5% of cases. Both *bored* and *sad* were classified as belonging to the *bored/sad* category in 80% of cases respectively; the remaining 20% were confused primarily with *neutral. Surprised* was correctly classified in 80% of cases.

The high recognition rates obtained suggest that regression trees are capable of representing the differences that occur in the source, as captured by LF parameters, between a *neutral* utterance and certain categories of emotional colouring, primarily capturing differences in activation level.

It is apparent from the formal listening tests that important emotion related voice information is lost in the transformations performed on the voice source parameters resulting in loss of differentiation between *angry* and *happy* and *sad* and *bored*. For example, comparing original and transformed parameter tracks for F0 in Fig. 3, it is clear that the transformations have yielded identical F0 parameter tracks for *sad* and *bored*. Transformation of *neutral* F0 into *angry, happy*, and *surprised* has resulted in identical F0 tracks for *happy* and *angry*, and *surprised* differs only towards the end of the utterance. Not all source parameters, however, fall together in the same way as F0. Compare, for example, the regression tree presented in Fig. 2, where for TL, *sad* and *angry* are each well separated from other emotion targets in the parameter transformation.

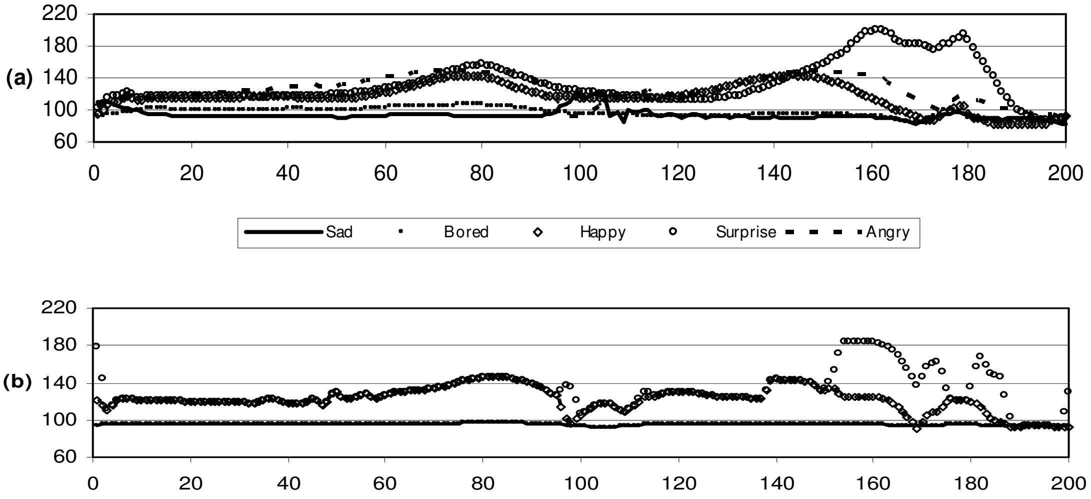

Figure 3: *(a) Original and (b) transformed F0 parameter tracks for five emotions.*

3. Discussion and conclusions

The analysis detailed in [1] suggests that each emotion is characterised by a specific combination of source parameter settings. Not only did the average source parameter values vary, but also the rate of variation of the parameters across an utterance was different. By using regression trees, we can map the changes that occur from an utterance encoding no emotion to an utterance in which emotion *has* been encoded.

The experiments above did not include changes in formants and bandwidths with changing emotions, but this is something that could also be investigated. It is however clear that the voice and its role in the encoding of emotion should not be ignored in emotion synthesis, and by the addition of voice source information to the rich knowledge of other acoustic correlates of emotion, better quality emotion synthesis may be achieved.

In an effort to improve the transformations we have recently enhanced the coding of syllable prominence in building and use of the regression trees, to include more detailed prosodic tags related not only to the syllable prominence (stressed/unstressed) but also to its role in the prosodic organisation of the intonational phrase, such as pre-head, head, nucleus, and tail. Statistical analysis of these new trees suggests that adding more detailed prosodic tags will improve the accuracy of transformed source parameter trajectories.

Moreover, utterance duration and speech rate information could be included. For example, in the experiments above, all utterance durations were scaled using anchor points and linear interpolation to the utterance duration of *neutral*. Future experiments should also manipulate durations and speech rate of the neutral utterance, as they are well-known correlates of emotion.

Such transformations have many possible implications for future synthesis use. For example, in applications directed at children or at learning environments, where standard audio texts are available, one might wish to add more emotional colouring to liven them up in specific contexts.

4. Acknowledgements

The research was funded by the EU FP6 Network of Excellence HUMAINE and by the Irish Research Council for Science and Engineering Technology-EMBARK initiative.

5. References

[1] Yanushevskaya, I.; Tooher, M.; Gobl, C.; Ní Chasaide, A., 2007. Time- and amplitude-based voice source correlates of emotional portrayals. *Proc. Affective Computing and Intelligent Interaction, ACII-2007.* Lisbon, 159-170.

[2] Juslin, P.; Scherer, K.R., 2005. Vocal expression of affect. In Harrigan, J., Rosenthal, R., Scherer, K.R. (eds.) *The New Handbook Methods in Nonverbal Behaviour Research*. Oxford: Oxford University Press, 65-135.

[3] Gobl, C.; Ní Chasaide, A., 2003. The role of the voice quality in communicating emotions, mood and attitude. *Speech Communication* 40, 189–212.

[4] Burkhardt, F.; Sendlmeier, W.F., 2000. Verification of acoustical correlates of emotional speech using formant-synthesis. *ISCA Workshop on Speech and Emotion.* Belfast, 151-156.

[5] Carlson, R.; Granström, B.; Nord, L., 1992. Experiments with emotive speech - acted utterances and synthesized replicas. *Proc. 2nd Int. Conf. Spoken Language Processing*, Alberta, Canada, 671-674.

[6] Audibert, N.; Vincent, D.; Aubergé, V.; Rosec, O., 2006. Expressive speech synthesis: evaluation of a voice quality centred coder on the different acoustic dimensions. *Proc. 3rd Int. Conf. on Speech Prosody,* Dresden, Germany.

[7] Fant, G.; Liljencrants, J.; Lin, Q., 1985. A four-parameter model of glottal flow. *STL-QPSR* 4/1985, 1-13.

[8] Klatt, D.H.; Klatt, L.C., 1990. Analysis, synthesis and perception of voice quality variations among female and male talkers. *JASA*, 87(2), 820- 857.

[9] Loh, W., 2002. Regression trees with unbiased variable selection and interaction detection. *Statistica Sinica*, 12, 361-386.