Voice Quality Interpolation for Emotional Text-To-Speech Synthesis

Oytun Turk[1,2] *Marc Schröder*[3] *Baris Bozkurt*[4] *Levent M. Arslan*[1,2]

[1] R&D Dept., Sestek Inc., Istanbul, Turkey
[2] Electrical and Electronics Eng. Dept., Bogazici University, Istanbul, Turkey
[3] DFKI GmbH Language Technology Lab, Saarbrücken, Germany
[4] TCTS Lab, Faculté Polytechnique de Mons, Mons, Belgium
oytun@sestek.com.tr schroed@dfki.de bozkurt@tcts.fpms.ac.be arslanle@boun.edu.tr

Abstract

Synthesizing desired emotions using concatenative algorithms relies on collection of large databases. This paper focuses on the development and assessment of a simple algorithm to interpolate the intended vocal effort in existing databases in order to create new databases with intermediate levels of vocal effort. Three diphone databases in German with soft, modal, and loud voice qualities are processed with a spectral interpolation algorithm. A listening test is performed to evaluate the intended vocal effort in the original databases as well as the interpolated ones. The results show that the interpolation algorithm can create the intended intermediate levels of vocal effort given the original databases independent of the language background of the subjects.

1. Introduction

The synthesis of expressive and emotional speech focuses on the generation of speech from text with desired emotions and relies on two kinds of issues: First, it is important to model the links between perceived emotions and the acoustic parameter configurations that cause this perception. This knowledge can be gathered through corpus analyses [1] and controlled listening experiments [2]. Second, acoustic parameters must be controllable so that the expressive prosody rules gathered in the first step can be implemented. This paper addresses a new method for improving the second issue, by providing a new means of controlling voice quality in concatenative synthesis, based on a combination of data recording and voice conversion techniques.

For concatenative emotional speech synthesis, the key issues are: how to record large speech databases with high voice quality and prosodic coverage and how to use these speech units to form target utterances with target voice quality and prosodic features without discontinuities. The TTS system design process is nearly impractical to tackle: recording of large databases with high phonetic, prosodic and voice quality coverage, tagging them phonetically, analyzing the signals to access acoustic parameters of voice quality and other prosodic information, and further use of this data in the unit-selection process.

The voice quality dimension of the presented problem is already very complex and the current state of the art algorithms for voice quality analysis/modification /synthesis is far from being high quality. This study targets simplification of the voice quality problem by applying voice conversion techniques to achieve some voice quality modification by reducing database size to be recorded, and simplifying the voice quality labeling and selection processes since only a certain group of voice quality speech will be needed for analysis. For demonstration of the proposed technique, we use diphone-based synthesis with varying voice qualities.

Until very recently, it has been impossible to control voice quality in diphone synthesis. Schröder and Grice have made a first step towards a limited control of voice quality [3]: For a male and a female voice, they recorded a full German diphone set with three levels of vocal effort: low ("soft" voice), medium ("modal" voice), and high vocal effort ("loud" voice). In a perception test with their male voice database, they showed that effort was perceived as intended despite the distortions to the original recordings introduced by the MBROLA algorithm.

The present paper addresses three issues: First, we want to verify that the recorded effort is perceived as intended for the female voice recorded by Schröder and Grice [3]. Second, we apply a spectral interpolation algorithm to interpolate between the recorded levels of vocal effort (soft, modal and loud), with the intention to create intermediate levels of effort. We construct new databases and test whether the TTS output obtained by using these databases are also perceived as intended. Third, we address the question whether the perception of effort from these recordings depends on the language background of the listeners.

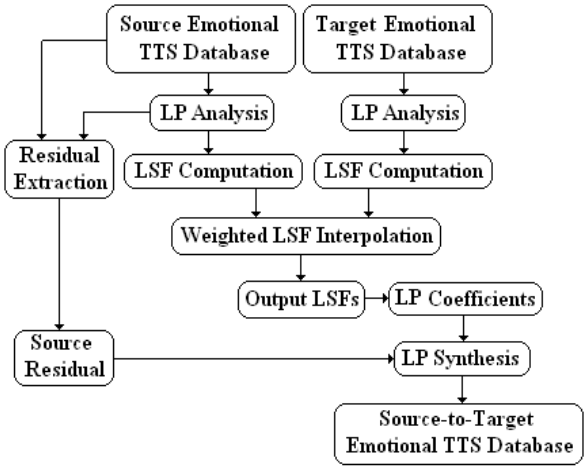

Figure 1: Flowchart of voice quality interpolation for emotional TTS.

Section 2 presents a description of the algorithm for voice quality interpolation and the database processing stage. In Section 3, the subjective listening test is described and the results are evaluated. In Section 4, a discussion of the results is presented.

2. Methodology

In the following sub-sections, we describe the main components of the proposed algorithm. Given the three emotional TTS diphone databases at different voice qualities in German (soft, loud, and modal), four interpolation steps are performed separately to obtain soft-to-modal, modal-to-soft, modal-to-loud, and loud-to-modal databases. Therefore, we had seven emotional diphone databases to perform synthesis, three of which were original and four were generated using the proposed method. Figure 1 shows a flowchart of the system.

2.1. Voice Quality Interpolation

From a signal processing point of view, voice quality variations mainly correspond to variations in spectral tilt, in the relative amount of aperiodic components in speech, and in some spectral variations in the low frequency part of the spectrum (like variations in the glottal formant frequency, in the bandwidth of first formant, etc.) and voice quality modifications of speech can be performed directly in the spectral domain [4]. To achieve such modifications proposed by D'Alessandro et al, it is necessary to decompose the speech signal into periodic and aperiodic parts, modify the amplitude spectrum of the periodic part by adaptive filtering and re-combine with the aperiodic component with a new mixing level. This study follows a simplified version of the suggested approach for a special case where recordings at different levels of vocal effort are available: our aim is to achieve voice quality interpolation via spectral interpolation of the speech signals without decomposition/recomposition of periodic and aperiodic components. Interpolation is performed through linear prediction (LP) analysis/re-synthesis and we use the line spectral frequencies (LSFs) that are commonly employed in speech coding and voice conversion applications [5]. Such an interpolation successfully modifies the amplitude spectrum envelope without degrading speech quality. One important factor is the role of residual signal and to study its effect different versions of interpolation are created. For each pair of voice qualities, the interpolation factor is set to 0.5 (corresponding to averaging of spectral envelopes) and two interpolation results are obtained by using each of the residual signals without further modification/mixing. In the soft-modal interpolation examples, soft-to-modal refers to the version where residual of the soft database is utilized and modal-to-soft refers to the version where residual of the modal database is utilized for re-synthesis.

The interpolation algorithm aims to estimate a set of LSFs that will represent a smoothly interpolated version of the vocal tract spectrum of two speech frames that are identical in terms of phonetics but different in terms of voice quality. Let D_s and D_t correspond to identical diphones in the source and target databases respectively and let us assume that we want to generate a diphone with intermediate voice quality source-to-target using D_s and D_t. The diphone recordings are first analyzed and LP coefficients are computed on a frame-by-frame basis using a window size of 20 ms. and a skip-rate of 10 ms. where each frame is windowed by a Hamming window. The sampling rate was 22050 Hz and an LP order of 24 was employed. The residual signal is extracted by inverse filtering the source signals with the LP filter frame-by-frame. LP coefficients are then converted to LSFs [6]. For each frame in D_s, the corresponding frame in D_t is found using the phonetic labels of diphones that are available in the databases using the linear mapping formula:

$$i_t = \left\lfloor \frac{(i_s - b_s)}{(e_s - b_s)}(e_t - b_t) + b_t + 0.5 \right\rfloor \qquad (1)$$

where i_s (i_t) is the index of the current speech frame in source (target) diphone D_s (D_t), b_s (b_t) and e_s (e_t) are the indices of the first and last speech frames in the phonetic label of current speech frame in D_s (D_t). The interpolated LSF vector, L_o, is obtained as a weighted average of L_s, the LSF vector for speech frame i_s in the source diphone D_s, and L_t, the LSF vector for speech frame i_t in the target diphone D_t, by:

$$L_o = (1 - r).L_s + r.L_t, \ 0 \le r \le 1 \qquad (2)$$

where L_o is the output LSF vector and r is the mixing ratio. We have used r=0.5 in the evaluations. L_o is converted to LP coefficients and the output diphone is synthesized using the source residual signal and the output LP coefficients using the overlap-add method. Figure 2a shows the LP spectra for the phoneme boundary frame (mid-frame) in diphone /aI-aI/ from the three original databases. In Figure 2b and 2c, LP spectra in original diphones and interpolated versions are shown.

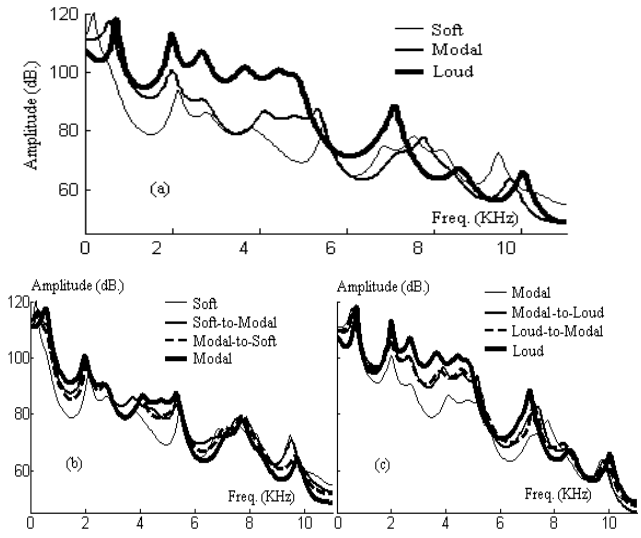

Figure 2: (a) Linear prediction (LP) spectra of the mid-frame of /aI-aI/ diphone in soft, modal, and loud databases, (b) LP spectra of the mid-frame in soft, modal, soft-to-modal, and modal-to-soft databases, (c) LP spectra of the mid-frame in modal, loud, modal-to-loud, and loud-to-modal databases.

2.2. Database Pre-processing for the Concatenator

As the concatenation algorithm, we use the well-known diphone-based synthesizer MBROLA (Multi-band re-synthesis overlap add) [7]. The MBROLA system performs an off-line pre-processing of the database, which facilitates high quality concatenation during run-time. The pre-processor re-

synthesizes short-time speech frames at a constant (average) pitch by re-sampling the amplitude spectrum and imposing constant phase envelope for low frequency part of the spectrum for all voiced frames. Such operations introduce some modification in the phase component and the spectral details of the spectrum (leading to some distortion in the periodic-aperiodic component ratio), but the spectral amplitude envelope is well preserved. The level of distortion in periodic-aperiodic ratio depends on many factors like actual periodic-aperiodic ratio of the particular signal, but most of the time it is at an inaudible level, which is also the case for our female voice databases.

3. Evaluations

We evaluated the perceptual properties of the recorded and interpolated databases using a listening test. As stimulus material, we designed four pseudo-sentences with the intention to simulate a plausible foreign language. In designing them, we took into account aspects of widely used phoneme inventory and syllable structure[1]. The pseudo-sentences, for which no "translation" was given, are transcribed below, using the SAM phonetic alphabet [8] and a GToBI description of the intonation contour [9]:

1. [ne: 'ka:l ?i:-'ba:m 'zu:t mo:-'le:n]
   L+H* L+H* H* H* L-%
2. ['kUn ze 'mi:-na lo:t be:-'la:m]
   L* H* L* H-^H%
3. [su: le-'mi:-na: fa-si: mo: 'le:-n@]
   L+H* H-% H* L-%
4. [me: 'lo:-ni: su 'na:-mo:]
   H* L* H-^H%

It can be seen that sentences 1 and 2 are more complex than sentences 3 and 4 with respect to their syllable structure and phoneme inventory: 1 and 2 contain closed syllables made of vowels, nasals, laterals, fricatives, and plosives, while 3 and 4 only contain open syllables made of vowels, nasals, laterals, and fricatives (but no plosives). Sentences 1 and 3 were generated with a final fall, suggesting a statement intonation, while sentences 2 and 4 were produced with a final rise, mimicking a question intonation.

Using the MaryXML representation language [10], we entered the transcriptions for these four sentences into the MARY TTS system, and converted them into the input format for the MBROLA synthesis algorithm using the standard German processing modules. Using the three original and the four interpolated German diphone databases, each sentence was then converted into seven sound files. In order to eliminate playback volume as a criterion in the perception test, the sound files were power-normalized, resulting in four sets of seven stimuli differing only in intended vocal effort.

We used the open-source tool RatingTest [11] for carrying out the listening tests. The seven versions of a sentence were presented as icons that can be clicked and dragged on a computer screen. The order in which the sentences were presented as well as the individual versions of a sentence were automatically randomized. Subjects heard the stimuli over headphones when they double-clicked an icon; they could then drag the icon to the adequate location on a scale indicating degree of effort perceived. They could repeatedly listen to the different versions of a sentence, compare them, and adjust their locations on the scale, until they were satisfied with the configuration. By clicking a confirmation button, they could then move on to rating the seven versions of the next sentence.

We carried out the test in Turkey, Germany, and Belgium. In order to ensure consistent instructions, we first formulated them in English, and then translated them into Turkish, German, and French, so that each subject could use the interface in their own native language. In each of the three countries, ten subjects (five male, five female; mean age 27.0 years; mostly students and laboratory staff) participated in the experiment.

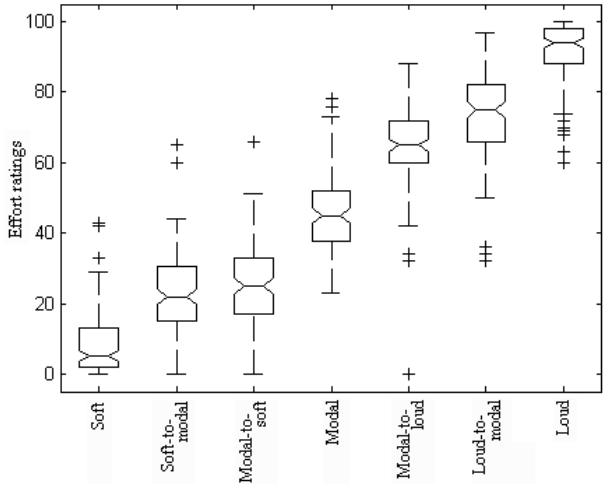

Figure 3: Effort ratings for each intended effort level group.

We performed a 4-way ANOVA of the results using SPSS, with perceived *effort* as the dependent variable, and the four independent variables *database* (representing the intended effort), *sentence*, *language* background of the listeners, and listener *gender*. By far the strongest main effect was that of *database* ($F(6, 672) = 997$, $p<0.001$), confirming very strongly our hypothesis that intended effort has an effect on perceived effort. The three other predictors also produced small but significant main effects: for *sentence*, $F(3,672)=3.2$, $p=0.02$; for *language*, $F(2,672)=4.7$, $p=0.01$; and for *gender*, $F(1,672)=5.5$, $p=0.01$.

The main effect of *language* is relevant for our research question (iii). Despite that small effect, ratings for different language groups look very similar (See Figure 4, left).

Similarly, the small but significant main effect of *sentence* does not seem to be linked to a systematic difference between the four sentences (See Figure 4, right).

In view of our research questions (i) and (ii), we investigated which of the databases were perceived in a significantly different way. The box plot (Figure 3) illustrates this: The interquartile ranges for most database ratings do not overlap, with the exception of the interpolated databases soft-to-modal and modal-to-soft (which overlap nearly completely) and modal-to-loud and loud-to-modal (which overlap to some extent). When we investigated pair-wise group means by 1-way ANOVA, perceived effort ratings for soft-to-modal and

[1] Thanks to William Barry and Martine Grice for very helpful suggestions on this topic. Two of the pseudo-sentences have been created in cooperation with University of Geneva (Klaus Scherer and Tanja Bänziger), who used them in recording a database of emotional speech suitable for cross-cultural studies.

modal-to-soft were not found to be statistically different (p=.07). Interestingly enough the modal-to-loud and loud-to-modal ratings showed differences that are statistically significant (p<.001). This result indicates that there might be a non-linear relationship for the perception of interpolated modes and/or the origin of the residual signal may explain the difference. However, this requires further research and more detailed tests.

We also investigated the effect of language, gender, and sentence on the effort ratings using a 1-way ANOVA test between soft-to-modal and modal-to-soft pair. None of these factors had significant effect on the effort ratings. We repeated this analysis for modal-to-loud and loud-to-modal pair and obtained the same result.

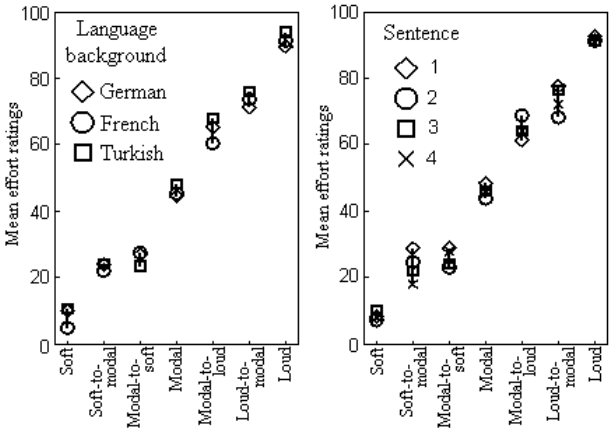

Figure 4: Mean effort ratings per database, for language background of the subjects (left) and sentences (right).

4. Discussion and Conclusions

The results of the perception test give the following answers to our research questions:

- The three recorded databases soft, modal, and loud are perceived as intended; indeed, they are very clearly distinguished.
- The interpolated databases are also perceived as intended, i.e. as intermediate levels of effort. Interpolation of the soft and the modal databases yielded perceptually equivalent results when a mixing ratio of r=0.5 is used. However, when interpolating between the modal and loud databases with r=0.5, the loud-to-modal database obtained by filtering the "loud" residual with an interpolated version of "loud" and "modal" filters resulted in higher perceived effort than the modal-to-loud database which is obtained by filtering the "modal" residual with an interpolated version of "modal" and "loud" filters. This result indicates that the residual plays a role in the perceived voice effort and its contribution needs to be studied. Theoretically an interpolation of residuals is also necessary however such modifications often result in audible distortions in the signal and was therefore not included in our algorithm. It is safer for the sake of cleanness of speech to adjust the mixing ratio appropriately.
- The effect of language background on the effort ratings is very small. Despite minor differences found in the mean effort ratings of French and Turkish listeners, the overall pattern in all three language groups is very similar.

In summary, we have shown that both recorded and interpolated databases are perceived as intended, quite independently of the language background of listeners and the phonetic string produced. Further research is necessary regarding the respective roles of the residual and the mixing ratio for interpolated voice quality.

5. Acknowledgements

The authors would like to thank to everyone who volunteered for the subjective listening test. Part of this research is supported by the EC Project HUMAINE (IST-507422), by EC Project SIMILAR (IST-507609), by Region Wallonne, Belgium, grant FIRST EUROPE #215095, and by Tubitak, Turkey, TIDEB 3020039.

6. References

[1] Schröder, M. Speech and Emotion Research: An overview of research frameworks and a dimensional approach to emotional speech synthesis. *PhD thesis, PHONUS 7, Research Report of the Institute of Phonetics, Saarland University,* 2004.

[2] Burkhardt, F. and Sendlmeier, W. F., "Verification of acoustical correlates of emotional speech using formant synthesis", in *Proc. of the ISCA Workshop on Speech and Emotion*, pp. 151-156, Northern Ireland, 2001.

[3] Schröder, M., and Grice, M., "Expressing vocal effort in concatenative synthesis", in *Proc. 15th International Conference of Phonetic Sciences*, Barcelona, Spain, pp. 2589-2592, 2003.

[4] D'Alessandro, C., and Doval, B., "Experiments in Voice Quality Modification of Natural Speech Signals: The Spectral Approach", in *Third ESCA/COCOSDA Workshop on Speech Synthesis 1998*, pp. 277-282.

[5] Huang, X., Acero, A., and Hon, H.-W. Spoken Language Processing – A Guide to Theory, Algorithm, and System Development, New Jersey, Prentice-Hall, Inc., 2001.

[6] Rothweiler, J., 1999, "A Root-finding Algorithm for Line Spectral Frequencies", in *Proc. of the IEEE ICASSP,* Phoenix, AZ, USA, vol. II, pp. 661-664, 1999.

[7] Dutoit, T. and H.Leich, "Text-to-speech synthesis based on a MBE re-synthesis of segments database"', *Speech Commun.,* Vol.13, p 435-440,1993.

[8] Wells, J.C., "SAMPA computer readable phonetic alphabet". In Gibbon, D., Moore, R. and Winski, R. (eds.), *Handbook of Standards and Resources for Spoken Language Systems*, Berlin and New York: Mouton de Gruyter, Part IV, Section B, 1997.

[9] Grice, M., Baumann, S., and Benzmüller, R. German intonation in autosegmental-metrical phonology. In Jun, S.-A., editor, Prosodic Typology, Oxford University Press, 2002.

[10] Schröder, M., and Breuer, S., "XML Representation Languages as a Way of Interconnecting TTS Modules", in *Proc.of the ICSLP,* Jeju, Korea, 2004.

[11] Schröder, M. RatingTest – Java software for designing and carrying out listening tests. http://ratingtest.sourceforge.net, 2005.