Interpolating expressions in unit selection

Marc Schröder**, DFKI GmbH, Saarbrücken, Germany

In expressive speech synthesis, a key challenge is the generation of flexibly varying expressive tone while maintaining the high quality achieved with unit selection speech synthesis methods. Existing approaches have either concentrated on achieving high synthesis quality with no flexibility, or they have aimed at parametric models, requiring the use of parametric synthesis technologies such as diphone, formant or HMM-based synthesis.

This extended abstract reports on on-going work exploring the addition of a certain degree of control over expressivity in a unit selection context. Rather than merely choosing *one* unit selection voice database in order to determine the expression contained in the generated speech, we use technology from the voice conversion domain to flexibly interpolate between *two* voice databases. This provides us with the possibility to generate a *continuum* of expressive tones between the two extremes defined by the two voice databases.

Spectral interpolation algorithm

The spectral interpolation method employed here has previously been used to interpolate diphone voices with different vocal effort [1]. The method is based on a linear predictive coding (LPC) paradigm of speech representation, using line spectral frequencies (LSFs) as a representation of LPC coefficients with good interpolation properties [2].

The method works as follows. Two utterances with the same phoneme chain are mapped to each other on the time axis. For each analysis frame in a given phoneme in the source utterance, the corresponding frame in the target utterance is determined by linearly scaling the phoneme durations. Analysis frames can be either pitch-synchronous or at a fixed frame rate. Both frames are represented as LSFs plus residual. Interpolation between these two frames is performed individually for each LSF. With LPC prediction order p, let lsf_m^S be the m-th LSF of the source frame, and lsf_m^T the m-th LSF of the target frame, for $1 \leq m \leq p$. Then we compute the interpolated output LSF

$$lsf_m^O = (1 - r) \cdot lsf_m^S + r \cdot lsf_m^T \tag{1}$$

where r is the mixing ratio, $0 \leq r \leq 1$. Higher r means a larger contribution of the target signal in the interpolated spectrum.

Before re-synthesising audio with the interpolated LPC filter, we need to scale the source residual with a gain factor computed by interpolating gain in the energy domain. In LPC analysis, the *prediction gain* is defined as the square root of the total energy of the prediction error, i.e. of the residual. From the gain of the source and target frames, g^S and g^T, the gain factor is computed as

$$gainfactor = (1/g^S)\sqrt{(1 - r) \cdot (g^S)^2 + r \cdot (g^T)^2} \tag{2}$$

** This work was supported by the projects HUMAINE (IST-507422) and PAVOQUE.

One frame of audio is resynthesised by filtering the gain-corrected residual of the source with the LPC filter defined by the interpolated LSFs lsf_m^O. Frames are combined into the resulting audio stream using a standard overlap-add mechanism.

Application in unit selection synthesis

We have integrated the interpolation algorithm into our unit selection speech synthesis platform MARY (`http://mary.dfki.de`), in a way that makes it easy to use the interpolation from markup.

Two unit selection voices can be interpolated by writing as input markup: `<voice name="voice1 with XY% voice2">`, where voice1 and voice2 are existing unit selection voices, and XY is a number between 0 and 100, indicating the relative weight of the spectrum from voice2 to be used in the interpolation. Thus, `"voice1 with 0% voice2"` corresponds to the original voice1, whereas `"voice1 with 100% voice2"` is a combination of the LPC residual from voice1 with the spectral envelope from voice2.

The algorithm first selects and concatenates units for each of the two voices separately; in the subsequent interpolation step, the unit durations serve as phoneme labels for the frame mapping.

We have tested the algorithm using two limited domain voices from the same speaker, generating "neutral" and "excited" soccer announcements. The "neutral" voice states the results in a rather matter-of-fact tone; the "excited" voice resembles the style of announcements made in soccer stadiums: high pitch, high vocal effort, and a relatively fast speech rate.

First informal listening tests confirm that the synthetic utterances generated with this interpolation algorithm are of good quality, with gradually changing spectral characteristics as the interpolation weight is changed. Noticeable distortions could be heard under two circumstances. Noise-like sounds were generated for some plosives when merging the excited spectrum into the neutral voice, probably due to different timing of silence vs. burst within the plosive units. This could be avoided by analysing the substructure of plosives with respect to acoustically similar sections, which would allow for a more appropriate time alignment. A weaker but noticeable type of distortion occurred when merging the neutral spectrum into the excited voice: at mixing ratios around 50%, some vowels were accompanied by a faint buzz noise. Despite these minor distortions, however, the overall degradation to intelligibility and naturalness seems very limited, and the interpolated voice exhibits vocal characteristics between the two original synthetic voices.

References

1. Turk, O., Schröder, M., Bozkurt, B., Arslan, L.: Voice quality interpolation for emotional text-to-speech synthesis. In: Proc. Interspeech 2005, Lisbon, Portugal (2005) 797–800
2. Paliwal, K.: Interpolation properties of linear prediction parametric representations. In: Proc. Eurospeech'95, Madrid, Spain (1995) 1029–1032