Affect Editing in Speech

Tal Sobol Shikler and Peter Robinson

Computer Laboratory, University of Cambridge, Cambridge, UK

Abstract. In this paper we present an affect editor for speech. The affect editor is a tool that encompasses various editing techniques for expressions in speech. It can be used for both natural and synthesized speech. We present a technique that uses a natural expression in one utterance by a particular speaker to other utterances by the same speaker or by other speakers. Natural new expressions are created without affecting the voice quality.

1 Introduction

Editing affect in speech has many desirable applications. Editing tools have become standard in computer graphics and vision, but speech technologies still lack simple transformations to manipulate expression of natural and synthesized speech. Such editing tools are relevant for the movies and games industries, for feedback and therapeutic applications, and more.

There is a substantial body of work in affective speech synthesis, see for example, the review by Schröder [6]. Morphing of affect in speech, meaning regenerating a signal by interpolation of auditory features between two samples, was presented by Kawahara and Matsui [3]. This work explored transitions between two utterances with different expressions in the time-frequency domain. Further results on morphing speech for voice changes in singing were presented by Pfitzinger [5], who also reviews other morphing related work and techniques.

However most of the studies explored just a few extreme expressions, and not nuances or subtle expressions. The methods that use prosody characteristics consider global definition, and only few integrated the linguistic prosody categorizations like f_0 contours [1][4]. The morphing examples are of very short utterances (one short word each), and few extreme acted expressions. None of these techniques leads to editing tools for general use.

In this paper we suggest an editing tool for affect in speech. We describe its architecture and a possible implementation which is based on known processing techniques. We also suggest a set of transformations of f_0 contours, energy, duration and spectral content, for the manipulation of affect in speech signals. This set includes operations like selective extension, shrinking, and actions like 'cut and paste'. In particular, we demonstrate how a natural expression in one utterance by a particular speaker can be transformed to other utterances, by the same speaker or by other speakers. The basic set of editing operators can be extended gradually to encompass a larger variety of transformations and effects. In the following sections we outline the method, show examples of subtle

J. Tao, T. Tan, and R.W. Picard (Eds.): ACII 2005, LNCS 3784, pp. 411–418, 2005.

expression editing of one speaker, demonstrate simple manipulations, and apply a transformation of an expression using another speaker's speech. We present here initial results. The techniques and their implementation still require refining and further validation with users and listeners.

2 Affect Editor

The affect editor, shown schematically in Figure 1, takes an input speech signal X, and allows the user to modify its conveyed expression, in order to produce an output signal $\tilde{X}$, with a new expression. The expression can be of an emotion, mental state or attitude. The modification can be a nuance, or might be a radical change. The operators that affect the modifications are set by the user.

The editing operators are derived in advance by analysis of an affective speech corpus. They can include a corpus of pattern samples for concatenation, or target samples for morphing. A complete system may allow the user to choose either a desired target expression that will be automatically translated into operators and contours, or choose the operators and manipulations manually. The editing tool should offer a variety of editing operators, such as changing the intonation, speech rate, the energy in different frequency bands and time frames, or add special effects.

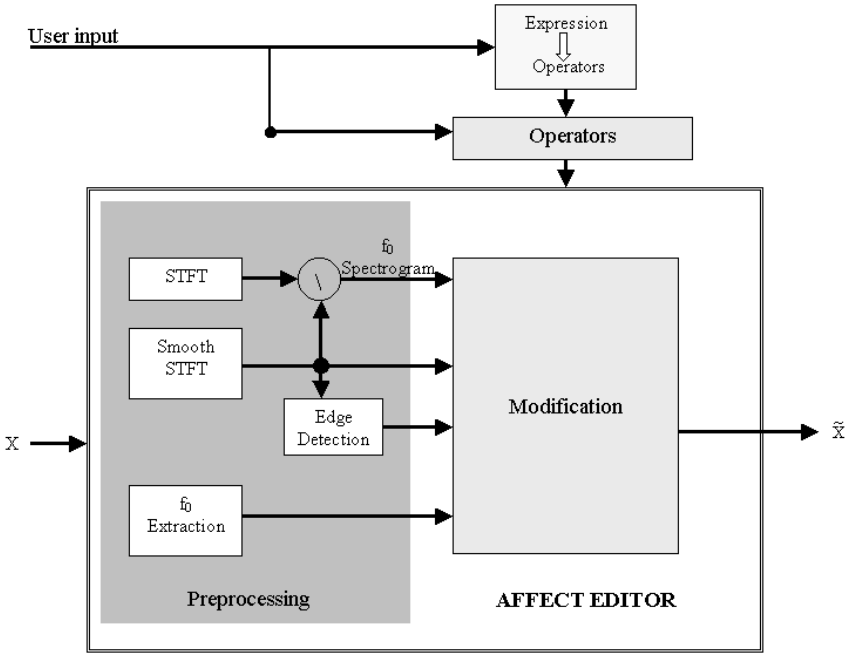

Fig. 1. A schematic description of an affect editing system

Extensions of this system rely heavily on an expressive inference system that should supply reliable operators and transformations between expressions and the related operators. Other extensions may include a graphical user interface that allows navigation among expressions and gradual transformations in time.

3 Implementation

The editor requires a preprocessing stage before editing an utterance. Post-processing is also necessary for reproducing a new speech signal. The input signal is preprocessed in a way that allows processing of different features separately. The method we use for preprocessing and reconstruction was described by Slaney [[7]], who used it for speech morphing. It is also close to Kawahara's method for morphing of affect. It is based on analysis in the time-frequency domain. The time-frequency domain is used because it allows for local changes of limited durations, and of specific frequency bands. From human computer interaction point of view, it allows visualization of the changeable features, and gives the user graphical feedback for most operations. We also use a separate f_0 extraction algorithm, so a contour can be seen and edited. These features make it a helpful tool for the psycho-acoustic research of features' importance.

The pre-processing stages include:

1. Short Time Fourier Transform, to create a spectrogram.
2. Calculating the smooth spectrogram using Mel-Frequency Cepstral Coefficients (MFCC). The coeffishcients are computed by resampling a conventional magnitude spectrogram to match critical bands as measured by auditory perception experiments. After computing logarithms of the filter-bank outputs, a low-dimensional cosine transform is computed. The MFCC representation is inverted to generate a smooth spectrogram for the sound which does not include pitch.
3. Divide the spectrogram by the smooth spectrogram, to create a spectrogram of f_0.
4. Extracting f_0. This stage simplifies the editing of f_0 contour.
5. Edge detection on the spectrogram, in order to find significant patterns and changes, and to define time and frequency pointers for changes. Edge detection can also be done manually by the user.

The pre-processing stage prepares the data for editing by the user. The affect editing tool allows editing of f_0 contour, spectral content, duration, or energy. Different implementation technique can be used for each editing operation, for example:

1. Changing the intonation can be done both by mathematical operations, or by using concatenation. Another method for changing intonation is to borrow f_0 contours from different utterances of the same speaker and other speakers. The user may change the whole f_0 contour, or only parts of it.

2. Change the energy in different frequency ranges and time frames. The signal is divided into frequency bands that relate to the frequency response of the human ear. A smooth spectrogram which represents these bands is generated in the preprocessing stage. Changes can then be made in specific frequency bands and time frames, or over the whole signal.
3. Change the speech rate. Extend and shrink the duration of speech parts by increasing and decreasing the overlap between frames in the inverse short-time Fourier transform. This method is good for the voiced parts of the speech, where f_0 exists, and for silence. The unvoiced parts, where there is speech but no f_0 contour, can be extended by interpolation.

These changes can be done on parts of the signal or on all of it. As will be shown below, operations on the pitch spectrogram and on the smooth/spectral spectrogram are almost orthogonal in the following sense. If one modifies only one of the spectrograms and then calculate the other from the reconstructed signal it will have minimal or no variations compared to the one calculated from the original signal. The editing tool has built-in operators and recorded speech samples. The recorded sample are for borrowing expression parts, and for simplifying imitation of expressions.

After editing, the system has to reconstruct the speech signal. Post-processing includes:

1. Regeneration of the new full spectrogram by multiplying the modified pitch spectrogram with the modified smooth spectrogram.
2. Spectrogram inversion, as suggested by Griffin and Lim [[2]]. Spectrogram inversion is the most complicated and time-consuming stage of the post-processing. It is complicated because spectrograms are based on absolute values, and do not give any clue regarding the phase of the signal. The aim is to minimize the processing time in order to improve the usability, and to give direct feedback to the user.

4 Examples

In this section we show some of the editing operations, with a graphical presentation of the results. We wanted to check if an affect editor is feasible with the current technology. The goal was to check if we could get new speech signals, that sound natural and convey new or modified expressions, and to experiment with some of the operators. We examine basic forms of the main desired operations, including changing f_0 contour, changes of energy, spectral content, and speech rate. For our experiment we used utterances from the Doors database, which consists of recordings of 15 people speaking Hebrew. Each speaker was recorded uttering repeatedly the same two sentences during a computer game, with approximately a hundred iterations each. The game elicited natural expressions and subtle expressions. It also allows tracking of dynamic changes among consecutive utterances.

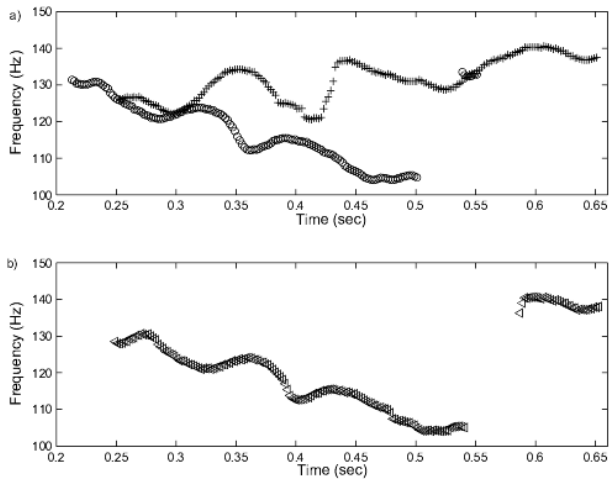

Fig. 2. Fundamental frequency (f_0) curves of 'sgor de-let'. a) original curves. The upper curve of 'uncertainty', the lower curve of 'determination' b) the curve of the edited signal, with combined pitch curve, and the energy and spectral content of 'uncertainty'.

Figure 2 presents features of the utterances 'sgor de-let' which means in Hebrew 'close door', uttered by a male speaker. Figure 2a represents the fundamental frequency curves of two original utterances. The higher curve shows the expression of uncertainty, and the lower curve shows determination. The uncertainty curve is long, high, and has a mildly ascending slope, while the determination curve is shorter and has a descending slope.

Figure 2b represents the curve of the edited utterance of uncertainty, with the combined f_0 curve generated from the two original curves, after reconstruction of the new edited signal. The first part of the original uncertainty curve, between 0.25sec and 0.55sec, was replaced by the contour from the determination curve. The location of the transformed part and its replacement were decided by using the extracted f_0 curves. The related parts from the f_0 spectrograms were replaced. A spectrogram of the new signal was generated by multiplying the new f_0 spectrogram by the original smoothed energy spectrogram. The combined spectrogram was then inverted. The energy and spectral content remained as in the original curve.

This manipulation yields a new natural speech signal, with a new expression, which is the intended result. We have intentionally chosen an extreme combination in order to check the validity of the editing concept, so the new expression is not necessarily identifiable. An end-user should be able to treat this procedure as the 'cut and paste', or 'insert from file' commands that are used by other types of editors. The user can use pre-recorded files, or to record the required expression to be used.

Figure 3 presents another set of operations on the utterance 'ptach de-let zo', which means 'open door this' (open this door) in Hebrew. First we manipulated

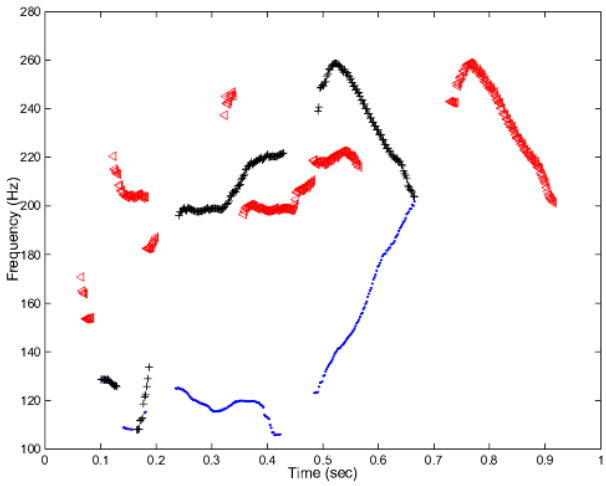

Fig. 3. f_0 contours of 'ptach delet zo' uttered by a female speaker (triangles), and a male speaker (dots), and the pitch of the edited male utterance (crosses)

local features of the fundamental frequency, as presented in Figure 3. We took an utterance by a male speaker, and replaces part of its f_0 contour with a contour of an utterance by a female speaker, with a different expression, using the same technique as in the previous example. The pitch of the reconstructed signal is shown in crosses. As can be seen, both the curve shape and its duration were changed. The duration was extended by inverting the original spectrogram with smaller overlap between frames. The sampling rate of the recorded signals was 32KHz, the short-time Fourier transform, and the f_0 extraction algorithm used frames of 50ms with original overlap of 48ms, which allow precision, calculation of low f_0, and flexibility of duration manipulations. After changing the intonation, we took the edited signal and changed its energy, by multiplying it by a Gaussian, so that the center of the utterance was multiplied by 1.2, and the sides, the beginning and the end of the utterance, were multiplied by 0.8. The new signal sounds natural, with the voice of the male speaker. The new expression is a combination of the two original expressions.

The goal here was to examine editing operators and get natural results. We employed a variety of manipulations, such as replacing parts of intonation contours with different contours by the same speaker and by another speaker, changing the speech rate, and changing energy by multiplying the whole utterance by a time dependent function. The results were new utterances, with new natural expressions, in the voice of the original speaker. These results were confirmed by initial evaluation with Hebrew speakers. The speaker was always recognized, and the voice sounded natural. On several occasions the new expression was perceived as unnatural for the specific person, or the speech rate too fast. It happened in utterances in which we had intentionally chosen slopes and f_0 ranges which are extreme for the edited voice. In some utterances the listeners

heard an echo. It occurred when the edges chosen for the manipulations were not precise.

Using pre-recorded intonation contours and borrowing contours from other speakers enable a wide range of manipulations of new speakers' voices, and may add expressions that are not part of a speaker's repertoire. A relatively small reference database of basic intonation curves can be used for different speakers. Time-related manipulations, such as extending the shrinking durations, and applying time dependent functions, extend the editing scope even farther. The system allows flexibility and a large variety of manipulations and transformations and yields natural speech.

Gathering these techniques and more, under one editing tool, and defining them as editing operators create a powerful tool for affect editing. However, for a full system, which is suitable for general use the algorithms should be refined, especially synchronization between the borrowed contours and the edited signal. Special consideration should be taken for the differences between voiced, where there is f_0, and unvoiced speech. Usability aspects should also be addressed, including processing time.

5 Summary

In this paper we have suggested the application of affect editing for non-verbal aspects of speech. Such an editor has many useful applications. We have demonstrated some of the capabilities of such a tool for editing expressions of emotion, mental state and attitudes, including nuances of expressions and subtle expressions.

We examined the concept using several operations,including borrowing f_0 contours from other speech signals uttered by the same speaker and by other speakers, changing speech rate, and changing energy in different time frames and frequency bands. We managed to reconstruct natural speech signals of the same speaker, with new expressions. These experiments demonstrate the capabilities of such an editing tool, although the details of the implementation should be refined. Future work should extend the scope of the operators and take into account usability issues including real-time processing.

Further extensions should include input from affects inference systems and labeled reference data for concatenation, automatic translation mechanism from expressions to operators, and a user interface that allows navigation among expressions.

References

1. Burkhardt F., Sendlmeier W. F.: Verification of Acoustical Correlates of Emotional Speech using Formant-Synthesis, ISCA Workshop on Speech & Emotion, Northern Ireland 2000, p. 151-156.
2. Griffin D. W., Lim J. S.: Signal Estimationm from Modified Short-Time Fourier Transform IEEE Trans. ASSP. **22** 1984 236-247.

3. Kawahara H., Matsui H.: Auditory Morphing Based on an Elastic Perceptual Distance Metric, in an Interference-Free Time-Frequency Representation, ICASSP'2003, pp.256-259, 2003.
4. Mozziconacci S. J. L., Hermes, D. J.: Role of intonation patterns in conveying emotion in speech, ICPhS 1999, p. 2001-2004.
5. Piftzinge H. R.: Unsupervised Speech Morphing between Utterances of any Speakers, Proceedings of the 10th Australian International Conference on Speech Science & Technology Macquarie University, Sydney, December, 2004. 545-550.
6. Schröder M.: Emotional speech synthesis: A review. In Proceedings of Eurospeech 2001, pages 561-564, Aalborg.
7. Slaney M., Covell M., Lassiter B.: Automatic Audio Morphing (ICASSP96), Atlanta, 1996, 1001-1004.