Inter-voice Audio Morphing

Jake Bouvrie, Tony Ezzat, and Tomaso Poggio

The Problem: This project seeks to develop a framework, which we have called "inter-voice morphing", for morphing between samples of speech that are identical in content, but produced by different speakers. Given two spoken phrases, we wish to smoothly morph between the characteristics that define the speakers in order to produce intermediate sequences that lie along the perceptual continuum connecting the two individuals.

Motivation: The theory and technology behind morphing between audio samples of different speakers has yet to solidify. However there are several applications for which such technology might prove useful. Speech synthesis and recognition applications requiring large template databases (e.g. concatenative speech synthesis) or, alternatively, the ability to smoothly interpolate within a small library of sequences, might find audio morphing particularly useful. In addition, behavioral experiments in psychology, psychophysics, and linguistics often times make use of large corpora of spoken samples possessing a range of desired spectral or perceptual properties. Such corpora are difficult and costly to produce; the automatic production of a speech database with user-tunable parameters from a small set of exemplars could prove to be an acceptable alternative to genuine human speech, or even open up new research opportunities in human perception.

Previous Work: Ezzat *et al.* [6] and Pfitzinger [4,5] have independently laid a foundation for speaker morphing, although using slightly different approaches. Ezzat borrows several key ideas originally introduced in the context of morphable models [1,2,3], to interpolate via a 1-D optical flow between the corresponding smooth spectra resulting from a cepstral decomposition of a pair of audio sequences. Pfitzinger utilizes a Linear Predictive Coding (LPC) based source-filter decomposition, but performs a similar DFW-based "warping" step to interpolate the LPC and residual spectrums. There has also been considerable work in the area of voice translation, where a mapping between a source voice and a destination voice is estimated from large audio databases. This research, however, is less relevant to the current project in that it does not allow for the smooth production of intermediate sequences, and also requires an entire dataset of speech examples for both speakers.

Approach: We approach the problem mainly from a signal processing perspective, in contrast to the learning or statistical approach, and do not require large databases of examples. We wish to morph two given audio sequences in the absence of any other information, including anchor points or other user-defined hints. We adopt a source-filter model of speech, which attempts to find a clean separation of the vocal-tract filter that shapes the glottal excitation signal from the excitation itself. Given a suitable decomposition of speech into smooth spectra and excitation signals, we interpolate smoothly between the respective smooth spectra and excitation components so as to construct intermediate sequences by traveling along an interpolation trajectory. The approach we plan to take follows from a combination of optical flow and dynamic programming, where one-dimensional correspondences along pairs of signals are computed (via DP) to produce an optimal path through the space of signals defined by the source and destination sequences. Once these correspondences have been computed, we can then morph between audio samples by setting a mixing parameter that defines the distance to travel along the lines connecting each set of corresponding points in the source signals, and then reconstructing the audio samples given the interpolated spectra and original phase information.

Progress: Using a corpus of recorded phrases, we have experimented with source-filter decompositions and subsequent interpolation steps. We have also examined hybrid time/frequency domain techniques that allow morphing of the spectral characteristics (in frequency), and morphing of the speaking rate as well as pitch-period shapes (in time). Finally, we continue to experiment with techniques [7,8] that allow for high-accuracy estimation of a time-domain signal given only the Fourier magnitude (a phase retrieval problem).

Future Work: We intend to focus specifically on the development of an algorithm for morphing the excitation (residual) component. While Ezzat *et al.* have shown that 1-D optical flow performs well for the smoothed spectrum component describing the vocal tract filter, excitation signals are inherently erratic and we expect that morphing this component of the speech decomposition will be significantly more challenging than interpolating the smoothed spectrum. Preliminary experimentation has also revealed that, as one might expect, we cannot simply cross-fade the residual component or reconstruct using the excitation signal from only one or the other of the source speech signals. Thus, accurate morphing of the pitch residuals is an important aspect of this project for future consideration. In a similar vein, the phase component of the spectral decomposition is a perceptually important factor, and must also be morphed or, alternatively, estimated from the interpolated spectral magnitude with high accuracy.

Research Support: Office of Naval Research (DARPA) Contract No. MDA972-04-1-0037, Office of Naval Research (DARPA) Contract No. N00014-02-1-0915, National Science Foundation-NIH (CRCNS) Contract No. EIA-0218506, and National Institutes of Health (Conte) Contract No. 1 P20 MH66239-01A1. Additional support was provided by: Central Research Institute of Electric Power Industry (CRIEPI), Daimler-Chrysler AG, Eastman Kodak Company, Honda Research Institute USA, Inc., Komatsu Ltd., Merrill-Lynch, NEC Fund, Oxygen, Siemens Corporate Research, Inc., Sony, Sumitomo Metal Industries, Toyota Motor Corporation, and the Eugene McDermott Foundation.

References:

[1] V. Blanz and T. Vetter. A morphable model for the synthesis of 3D faces. In Alyn Rockwood, editor, Proceedings of SIGGRAPH 2001, Computer Graphics Proceedings, Annual Conference Series, pages 187–194, Los Angeles, 1999. ACM, ACM Press / ACM SIGGRAPH.

[2] T.Ezzat, G.Geiger, and T.Poggio. Trainable videorealistic facial animation. In Proceedings of SIGGRAPH 2002, volume 21, pages 388–398, San Antonio, Texas, 2002.

[3] M. Jones and T. Poggio. Multidimensional Morphable Models: A Framework for Representing and Matching Object Classes. in Proceedings of the Sixth International Conference on Computer Vision, Bombay, India, 683-688, January 4-7, 1998.

[4] Pfitzinger, H.R. DFW-based Spectral Smoothing for Concatenative Speech Synthesis. In Proc. ICSLP 2004, vol. 2, pp. 1397-1400. Korea. Oct, 2004.

[5] Pfitzinger, H.R. Unsupervised Speech Morphing between Utterances of any Speakers. In Proc. of the 10th Australian Int. Conf. on Speech Science and Technology (SST 2004), pp. 545-550. Sydney. Dec., 2004.

[6] T.Ezzat, E.Meyers, J.Glass, and T.Poggio. Morphing Spectral Envelopes using Audio Flow. In Proc. ICASSP, Lisbon, Portugal, September, 2005.

[7] D.W. Griffin and J.S. Lim. Signal estimation from modified short-time fourier transform. IEEE Trans. Acoust., Speech, Signal Processing, vol. ASSP-32, no. 2, pp. 236--243, Apr 1984.

[8] M. Slaney, “Pattern Playback in the 90s”, In Proc. NIPS, 1994: 827-834.