Audio Morphing

Tony Ezzat

Artificial Intelligence Laboratory and
The Center for Biological and Computational Learning
Massachusetts Institute of Technology
Cambridge, Massachusetts 02139

http://www.ai.mit.edu

The Problem: In this work, we tackle the problem of morphing between different audio sequences. The system should take as input 2 audio sequences, and produce as output intermediate audio sequences that represent natural exemplars lying between the 2 input sequences.

Motivation: Audio morphing might have important applications in speech recognition, speech synthesis, music synthesis, and other applications where large corpora are recorded and there is a strong need to interpolate between the exemplars in the corpora to produce new exemplars.

Previous Work: There has been a spate of recent work on *voice conversion* [1, 4, 5], where a reference speaker speech sample is warped to match the statistical properties of a target speaker. Most authors resort to mixed time- and frequency- domain methods to alter pitch, duration, and spectral features.

Audio morphing [3] is closest in spirit to the goal of this work. The author used dynamic time-warping to time-align two speech samples, cross-faded the respective smoothed spectrograms, and warped a pitch residual to morph between two sounds.

Approach: In this work, we aim to explore and develop a purely time-domain method of audio morphing, motivated by the success of methods such as TD-PSOLA [2] for warping audio sequences.

Research Support: Research at CBCL is sponsored by grants from: Office of Naval Research (DARPA) under contract No. N00014-00-1-0907, National Science Foundation (ITR) under contract No. IIS-0085836, National Science Foundation (KDI) under contract No. DMS-9872936, and National Science Foundation under contract No. IIS-9800032 Additional support was provided by: Central Research Institute of Electric Power Industry, Center for e-Business (MIT), Eastman Kodak Company, DaimlerChrysler AG, Compaq, Honda R&D Co., Ltd., Komatsu Ltd., Merrill-Lynch, NEC Fund, Nippon Telegraph & Telephone, Siemens Corporate Research, Inc., Toyota Motor Corporation and The Whitaker Foundation.

References:

[1] A. Kain and M. Macon. Spectral voice conversion for text-to-speech synthesis. In *Proc. ICASSP*, pages 285–288, May 1998.

[2] E. Moulines and F. Charpentier. Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. *Speech Communication*, 9:453–467, 1990.

[3] M. Slaney, M. Covell, and B. Lassiter. Automatic audio morphing. In *Proc. ICASSP*, Atlanta, Georgia, 1996.

[4] Y. Stylianou, O Cappe, and E. Moulines. Statistical methods for voice quality transformation. In *Proc. Eurospeech*, pages 447–450, Madrid Spain, September 1995.

[5] H. Valbret, E. Moulines, and J.P. Tubach. Voice transformation using psola technique. *Speech Communication*, 11:175–187, 1992.