January 2011

M.Sc. Research Proposal
Submitted by Itai Katsir

Artificial Bandwidth Extension of Band-Limited Speech Based on Vocal-Tract Shape Estimation

Artificial Bandwidth Extension (ABWE) of speech is a technique for estimating missing frequency bands in band-limited speech signals. In traditional telephony networks, the speech is band limited to a narrowband (NB) frequency range of 300-3400 Hz. This band limitation causes degradation in speech quality and intelligibility. The aim of an ABWE algorithm is to improve NB speech quality by estimating the missing frequency bands from the wideband (WB) speech extending from 50-7000 Hz. The motivation for using this technique is to allow better speech quality for end users without a costly upgrade of the telephone network.

The common approach for ABWE relies on the source-filter model of speech production. It assumes that the estimation of missing frequency bands could be divided into two independent tasks of estimating the excitation source and the vocal tract filter. The current state of the art research in this field addresses two major challenges:

- Estimation of the WB vocal-tract filter – crucial for achieving high intelligibility speech.
- Estimation of missing-bands gains – crucial for high quality of estimated WB speech.

The common approach for tackling the first challenge is to map NB speech features to WB speech features using techniques like codebook mapping, linear mapping, and statistical approaches like GMM and HMM. The current techniques still face difficulties in the estimation of some speech sounds, especially unvoiced sounds. They also show large quality variation, which is speaker dependent. The second challenge is addressed by estimating the gains of missing bands directly from the NB speech features, or by gain adjustment of the estimated WB speech signal to match the gain of the received NB speech signal. Large estimation errors of missing-bands gains cause noticeable artifacts that result in speech quality degradation.

Our main research goals are:

- Improve estimated WB signal intelligibility by applying a phoneme-dependent vocal-tract filter estimation.
- Reduce quality variations obtained for different speakers, by speaker-dependent estimation of the vocal tract filter.
- Reduce speech artifacts by applying a better gain estimation of the reconstructed WB speech signal.

Accordingly, we propose a method for WB vocal-tract filter estimation which combines speech phonetic content estimation and speaker vocal-tract shape estimation. These two estimation tasks require choosing an appropriate statistical model and a statistical model training procedure to improve the estimation. They also require selection of an adequate feature set to represent the speech phonetic content and the vocal tract shape. We also propose to improve the common gain adjustment process by iteratively tuning the estimated vocal-tract shape.

Supervised by: Prof. David Malah and Prof. Israel Cohen