Pitch-synchronous Decomposition of Mixed-source Speech Signals

Philip J.B. Jackson and Christine H. Shadle

Dept. of Electronics and Computer Science, University of Southampton, Southampton SO17 1BJ, U.K.

Abstract: As part of a study on the acoustic mechanisms involved in aspiration, a method was developed for decomposing an acoustic signal into harmonic (voiced) and anharmonic (unvoiced) components, based on a hoarseness metric by H. Muta, T. Baer, K. W'agatsuma, T. Muraoka and H. Fukuda, 1988, "A pitch-synchronous analysis of hoarseness in running speech", J. Acoust. Soc. Am., 84(4):1292-1301. Their pitch-synchronous harmonic filter (PSHF) was extended (to EPSHF) to yield time series of both components. Results of applying the EPSHF to synthetic and recorded speech signals are given.

To decompose a speech waveform into harmonic and anharmonic components, which, ideally, would contain the vocal-tract-filtered voicing source, and all filtered noise sources, respectively, the pitch-synchronous harmonic filter, described by Muta et al. (1988) for quantifying hoarseness, was extended to yield reconstructed and improved time series estimates. The extracted 'voiced' and 'unvoiced' estimates enable both time- and frequency-domain analysis, e.g. for comparison of signal envelopes, synchrony of events, or identification of the location and type of source (via trough frequencies of anharmonic spectrum).

RECORDING AND ANALYSIS METHODS

An adult male subject (PJ), speaker of British English R.P., recorded a corpus consisting of repeated /pVFV/ nonsense words (V=/a,i,u/, F=/s,z/) in normal, breathy, stage-whispered and whispered voice in a sound-treated room. The sound pressure at 1 m was measured using B&K Type 4165 microphone and Type 2636 amplifier, using 20-20kHz band-pass, linear filter, and recorded onto DAT with a Sony TCD-D7, whose sampling frequency is 48 kHz. The 16-bit data were then transferred digitally to computer for analysis.

Ensemble averages were generated by marking equivalent locations in an array of tokens (e.g. release of a stop, or start/end of a fricative), and summing the sound power of the discrete Fourier transform (DFT) of the corresponding windowed portions from each token. Thus consistent features were amplified in relation to others, and also an indication of measurement variability was obtained.

Figure 1: Extended Pitch-Synchronous Harmonic Filter (EPSHF) architecture.

The pitch-period, \( N \) sample points, was computed according to Muta et al. (1988), which uses a von Hann-windowed section of speech signal, \( s_k \), four pitch periods long, \( k = \{0, 1, \cdots, 4N - 1\} \). The PSHF, \( H \) (see Fig.1), has coefficients \( h_n = \frac{S_n}{4} \) for \( n = \{4, 8, \cdots, 4(N - 1)\} \), otherwise \( h_n = 0 \), for frequencies at harmonics of the fundamental \( f_0 = \frac{500}{N} \), where \( f_s \) is the sampling frequency. The original estimate of the
‘voiced’ spectrum comprises a smeared version of these harmonic coefficients, $\hat{V}_{n-1} = 0.25h_n$, $\hat{V}_n = 0.5h_n$, $\hat{V}_{n+1} = 0.25h_n$, which is a frequency-domain equivalent of the windowing. The extended estimate of the ‘unvoiced’ spectrum was formed by interpolating the estimate $\hat{U}_n = S_n - \hat{V}_n$. The interpolation operator, $L$, has coefficients, $l_n = \{[\hat{U}_{n-1}]^2 + [\hat{U}_{n+1}]^2\}^{\frac{1}{2}}$, for $n = \{4, 8, \ldots, 4(N - 1)\}$. These coefficients were smeared, as before, and added to $\hat{U}_n$ to form $\hat{U}_n$, whose inverse DFT (IDFT) gives the anharmonic signal $\hat{u}_k$, which, when subtracted from $s_k$, leaves the harmonic component $\hat{v}_k$.

RESULTS

Synthetic signals, $s$, were processed to evaluate the performance of the EPSHF, and to identify any short-comings or artefacts. Tests were performed on three signals: $s_a = g + w$, $s_b = g + (w \times g)$, and $s_c = g \ast f_1 + (w \times g) \ast f_2$, where $g$ is periodic, $w$ is Gaussian, white noise, $f_1$ and $f_2$ represent filter impulse responses and $\ast$ denotes convolution. Figure 2a illustrates analysis of $s_c$, where the modulation of the filtered noise is preserved in the reconstructed signal, $\hat{u}$. In this example, the EPSHF improved the (a.c.) signal-to-noise ratio of the anharmonic part by 11.6 dB (from -5.9 dB to 5.7 dB).

Figure 2b shows how one token, a normally-phonated [pʰaɑ], was decomposed by the EPSHF. One can see in $\hat{u}$ a smooth and clean estimate of the quasi-periodic component. The anharmonic estimate, $\hat{u}$, contains most of the noisy speech components: the burst (20 ms), and initial frication and aspiration (30-70 ms), but also some noise during the vowels, which grows in [z] (330-420 ms) and at the cessation of voicing (720 ms). However, there are also glitches, a by-product of processing, at voice onset (70-100 ms) and other transient stages (200, 270, 450 ms) where there are either rapid changes in $f_0$ or voicing amplitude, or both. Interestingly, at the development of the fricative (310-360 ms), $\hat{u}$ exhibits modulation by $f_0$, which was typical of other tokens of [pʰaɑ]. Examples of other phonation modes (breathy, stage-whispered) for [pʰaɑ] were examined, and showed similar but exaggerated features, since the magnitude of the anharmonic components was greater, as was the degree of jitter and shimmer. Figure 3a shows ensemble-averaged spectra of [z] in [pʰaɑ] context, its harmonic and anharmonic estimates, and $\hat{u}$ in [pʰaɑ]. As expected, the anharmonic [z] spectrum is similar in shape and amplitude to [s]. Figure 3b shows a waterfall plot of ensemble-averaged spectra, from [pʰ] release to voice onset at approximately 5 ms intervals. The deep troughs in the burst spectrum (bottom curve) are indicative of a source localised at the lips. Just before voice onset (top), anti-resonances shift from 2.2 to 2.0 kHz, and formants rise (e.g. $F_3$, from 2.3 to 2.7 kHz).

CONCLUSION

The EPSHF plausibly decomposed breathy vowels, but the harmonic component of voiced fricatives still contained significant noise, whose ensemble-averaged spectrum was similar in shape to (though weaker than) the anharmonic one. The algorithm performed best on sustained sounds. Tracking errors at rapid transitions, and errors due to jitter and shimmer, were spuriously attributed to the anharmonic component. However, this component clearly revealed various features of the noise source, such as modulation in voiced fricatives. Further work will explore the interactions of phonation, breathy speech and turbulence noise, and the use of a reference signal, e.g. from a laryngograph, to enhance the filter performance.

Figure 3: Ensemble-averaged spectra of (a) mid-fricative in normally-phonated [paFa] context (F=/z,s/; 8 tokens, 85 ms window), with EPSHF harmonic and anharmonic decomposition of [z], and (b) release of [pʰ] to voice onset (16 tokens, 21 ms window), in [paFa] context, 10 dB between tick-marks and between each frame.