A PITCH ESTIMATION FILTER ROBUST TO HIGH LEVELS OF NOISE (PEFAC)

Sira Gonzalez and Mike Brookes

Imperial College London
Department of Electrical and Electronic Engineering
London SW7 2AZ, UK
email: {sira.gonzalez08, mike.brookes}@imperial.ac.uk

ABSTRACT

We present PEFAC, a fundamental frequency estimation algorithm that is able to identify the pitch of voiced frames reliably even at negative signal to noise ratios. The algorithm combines non-linear amplitude compression, to attenuate narrow-band noise components, with a comb-filter applied in the log-frequency power spectral domain, whose impulse response is chosen to attenuate smoothly varying noise components. We compare the performance of our algorithm with that of other widely used algorithms on a subset of the TIMIT database and demonstrate that it performs exceptionally well in both high and low levels of additive noise.

1. INTRODUCTION

The estimation of fundamental frequency, or pitch, is an essential component of many speech processing applications and numerous approaches have been described in the literature. Pitch estimators may be broadly divided into three groups according to whether they operate in the time, frequency or time-frequency domain. Typically, the first category finds peaks in the autocorrelation function, the second looks for harmonic peaks in the power spectrum while the third performs time-domain analysis on the outputs of a bank of bandpass filters. In many cases, an algorithm identifies multiple pitch candidates in each time frame and then uses temporal continuity constraints to select between them.

In situations where there is a high level of acoustic noise or where the distance between microphone and talker is large, the signal to noise ratio (SNR) of an acquired speech signal can be very poor. In such circumstances, the performance of pitch estimation algorithms degrades [1], and many methods become unusable below 0 dB SNR. In recent years a number of noise-robust algorithms have been designed but reliable fundamental frequency estimation at negative SNRs remains a challenging problem.

In this paper, we propose a new frequency-domain algorithm for pitch estimation that is robust to high levels of noise. Many frequency-domain algorithms begin by selecting isolated peaks in the short-time power spectrum, which are difficult to identify at poor SNRs, as potential pitch harmonics [2, 3]. However, in [4], instead of identifying isolated peaks, a comb-filter is used in the linear frequency domain to calculate a weighted sum of the harmonic amplitudes. For this method the fundamental frequency of the comb-filter, initially unknown, has to match the pitch. A sub-harmonic-summation (SHS) method in the log-frequency domain is proposed in [5], where the spectrum is shifted along the log-frequency axis, weighted and summed. Based on the same idea, [6] convolves the spectrum in the log-frequency domain with a train of delta functions harmonically spaced and selects the highest peak. Our algorithm, similarly, estimates the fundamental frequency of each frame by convolving its power spectral density in the log-frequency domain with a filter that sums the energy of the pitch harmonics while rejecting additive noise that has a smoothly varying power spectrum. Amplitude compression is applied before filtering to attenuate narrowband noise components.

2. PROPOSED METHOD

For a perfectly periodic source at frequency f_0, our signal model at time t in the power spectral density domain is

$$Y_t(f) = \sum_{k=1}^{K} a_{k,t}\delta(f - kf_0) + N_t(f) \tag{1}$$

where $N_t(f)$ represents the power spectral density of the unwanted noise and $a_{k,t}$ the power of the k^{th} harmonic. In the log-frequency domain, the signal model can be expressed as

$$Y_t(q) = \sum_{k=1}^{K} a_{k,t}\delta(q - \log k - \log f_0) + N_t(q) \tag{2}$$

where $q = \log f$. In this domain, the spacing of the harmonics is independent of f_0 and their energy can therefore be combined by convolving $Y_t(q)$ with a filter with impulse response

$$h(q) = \sum_{k=1}^{K} \delta(q - \log k) \tag{3}$$

The convolution $Y_t(q) * h(q)$ will include a peak at $q_0 = \log f_0$ and additional peaks corresponding to simple rational multiples of f_0.

2.1 Filter definition

In practice, the width of each harmonic peak will be broadened due to the analysis window and the rate of change of f_0. Accordingly we use a filter with broadened peaks having the impulse response

$$h(q) = \beta - \log(\gamma - \cos(2\pi e^q)) \tag{4}$$

for $\log(0.5) < q < \log(K+0.5)$ and $h(q) = 0$ otherwise. γ is an algorithm parameter that controls the peak width while β is chosen so that $\int h(q)dq = 0$. The number of peaks, K, is restricted to 10 in order to reduce the response of $Y_t(q) * h(q)$

at values of q corresponding to subharmonics of f_0. Fig. 1 shows $h(q)$ for $\gamma = 1.5$.

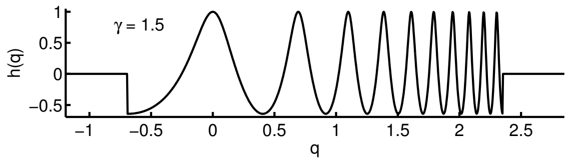


Figure 1: Filter $h(q)$ from (4) for $\gamma = 1.5$.

Because $h(q)$ is chosen to have zero mean, a white noise term, $N_t(q)$, will be suppressed by the filter. Moreover, because the peaks in $h(q)$ are each approximately symmetrical with zero mean, $N_t(q)$ will be suppressed if $\frac{dN_t(q)}{dz}$ is approximately constant over the range $\log(k-0.5) < q < \log(k+0.5)$ for each $0 < k \leq K$. In practice, this means that any smoothly varying noise power spectral density will be greatly attenuated by the filter.

2.2 Compression

Although, as we have seen, noise with a smoothly varying spectrum will be suppressed by the filter $h(q)$, some noise sources contain high amplitude narrowband components which may dominate the filter output. In order to avoid this, we apply compression to the spectrum of each time frame before convolving with $h(q)$ by setting

$$Y'_t(q) = Y_t(q)^{\alpha_t(q)} \tag{5}$$

where t is the time index. To determine the compression exponent, $\alpha_t(q)$, we first calculate the smoothed spectrum $\overline{Y}_t(q)$ by lowpass filtering $Y_t(q)$ in both time and log-frequency. In the absence of noise, we expect $\overline{Y}_t(q) \approx L(q)$, the long-term average spectrum of speech [7, 8]. Accordingly we normalize $\overline{Y}_t(q)$ and $Y_t(q)$ to the power of $L(q)$ and set the compression exponent to be

$$\alpha_t(q) = \frac{\log L(q)}{\log \overline{Y}_t(q)} \tag{6}$$

A strong narrowband noise source at q_n will result in $\overline{Y}_t(q_n) \gg L(q_n)$ and the resultant $\alpha_t(q_n) \ll 1$ will compress its amplitude. In addition, the power normalization of $\overline{Y}_t(q)$ means that noise free speech spectral components at other values of q will be enhanced because at these frequencies $\overline{Y}_t(q) < L(q)$.

2.3 Fundamental frequency estimation

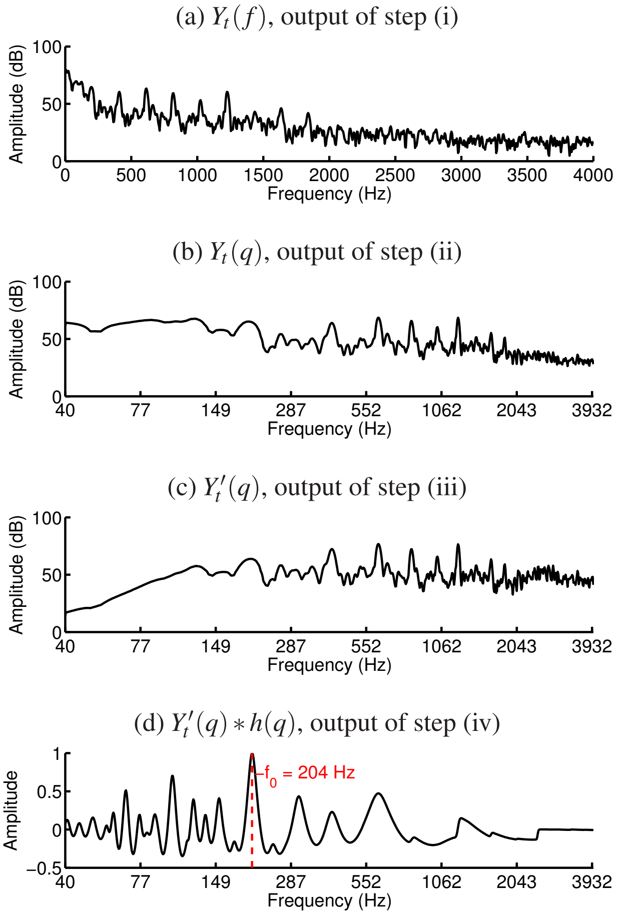
The complete PEFAC (Pitch Estimation Filter with Amplitude Compression) therefore comprises the following steps whose outputs are shown in Fig. 2 for a single voiced frame corrupted by car noise:

(i) transform the input signal to the time-frequency domain using the short-time Fourier transform (STFT), $Y_t(f)$,

(ii) interpolate the power spectral density (PSD) of each frame onto a log-spaced frequency grid, $Y_t(q)$,

(iii) find $\alpha_t(q)$ so that the normalized smoothed spectrum $\overline{Y}_t(q)$ equals $L(q)$ and calculate the compressed PSD, $Y'_t(q)$,

(iv) convolve the compressed PSD, $Y'_t(q)$, with the analysis filter, $h(q)$, and select the highest peak in the feasible range as the estimated pitch.


(a) $Y_t(f)$, output of step (i)

(b) $Y_t(q)$, output of step (ii)

(c) $Y'_t(q)$, output of step (iii)

(d) $Y'_t(q) * h(q)$, output of step (iv)

Figure 2: Algorithm processing steps for a single voiced frame of speech corrupted with car noise. (a) PSD in dB, (b) PSD in dB in a log-frequency grid, (c) compressed PSD in dB in a log-frequency grid, and (d) normalized output of the filter and fundamental frequency, f_0.

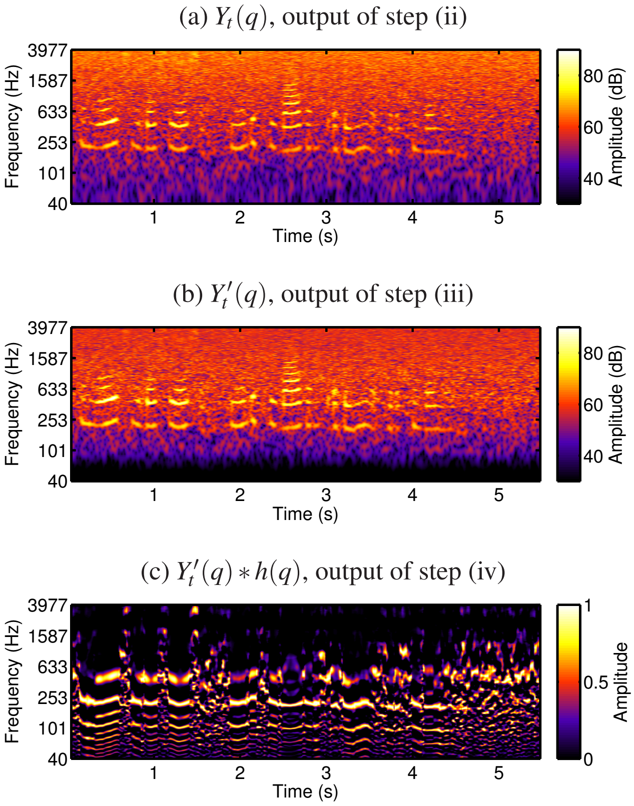
In Fig. 2 we see that the low frequency noise that masks the fundamental in (a) is greatly attenuated in (c) and that a clear peak at 204 Hz is visible in (d) despite being absent in the original spectrum.

The algorithm does not impose any temporal continuity constraints on the pitch estimates. Despite this, the algorithm results in very few gross pitch errors even at poor SNRs as demonstrated in Section 4.

3. EXPERIMENTS

The pitch estimator described above includes a number of algorithm parameters whose values were determined empirically using a development test set. The STFT used a Hamming analysis window of 90 ms duration; this is long enough to resolve the pitch harmonics even for low values of f_0 but short enough to limit the pitch variation within a frame. Each windowed input frame is zero-padded to 360 ms to aid the interpolation stage at low frequencies and the inter-frame time increment is 10 ms.

The spectrum of each frame is interpolated onto a



(a) $Y_t(q)$, output of step (ii)

(b) $Y_t'(q)$, output of step (iii)

(c) $Y_t'(q) * h(q)$, output of step (iv)

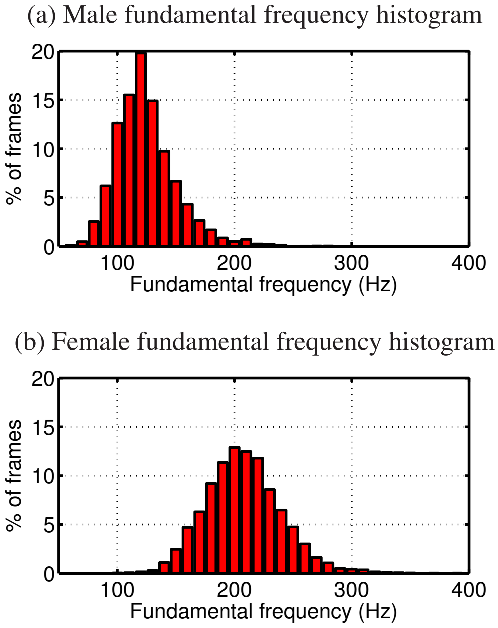
Figure 3: Output of various steps of PEFAC for a speech file corrupted with white noise at -5dB SNR. (a) Noisy speech log-frequency spectrogram, (b) compressed noisy speech spectrogram, and (c) output of the pitch analysis filter normalized at each frame to its peak value.

logarithmic grid ranging from 40Hz to 4kHz with a frequency resolution of 0.58%. Conceptually the sampled spectrum is first converted to a continuous spectrum using linear interpolation and this is then resampled using a variable width triangular sampling kernel. In practice the two stages are combined and the continuous spectrum is not calculated explicitly [8].

The smoothed spectrum $\overline{Y_t}(q)$ in the amplitude compression step is calculated using a uniform moving average filter with support $Q = 1.15$ in the log-frequency axis and averaging over the entire file (typically of 3-5s duration) in the time axis.

Following amplitude compression, the resampled spectrum of each frame is convolved with the filter $h(q)$ from (4). The optimum value of the parameter γ depends on the nature of the noise and the value 1.5 was chosen as the best compromise. For each frame, the position of the highest peak in the filtered output is selected as the estimated pitch.

The upper graph of Fig. 3 shows the spectrogram of a speech signal corrupted with white noise at -5dB SNR. The middle graph shows the effect of amplitude compression in which it can be seen that the noise has been significantly attenuated at low frequencies where little speech energy is present. The lower graph shows the output of the pitch estimation filter in which, for clarity, each frame has been normalized to its peak value. It can be seen that during voiced frames the filter output shows a strong peak at the correct pitch (about 200Hz) together with weaker peaks at rational multiples of this pitch.



(a) Male fundamental frequency histogram

(b) Female fundamental frequency histogram

Figure 4: Fundamental frequency distribution for males and females in the core test set from the TIMIT database

4. RESULTS

In this section, the performance of the proposed fundamental frequency estimator is evaluated. We used the core test set from the TIMIT database [9] which contains 16 male and 8 female speakers each reading 8 distinct sentences. Thus the core test material consist of 192 sentences containing a total of 28,473 voiced frames. The selection of this database was based on the wide range of accents and speakers present in it.

The ground truth for the fundamental frequency was determined using Praat [10] on clean speech. Errors in the estimation given by Praat were corrected manually. Fig. 4 shows the fundamental frequency distribution of males and females in the database.

Additive noise from the RSG-10 database [11] was added to the speech files to generate the noisy test signals. Three types of noise were used at SNRs from -20 to $+20$dB: white noise, car noise and babble. The measurement of SNR used ITU-T P.56 [12, 8] for the speech level and unweighted power for the noise.

For performance comparison, RAPT [13, 8], YIN [14] and Jin & Wang (J&W) [15, 8] were used. The first two of these are time-domain algorithms while the third is a time-frequency algorithm. The J&W algorithm was modified to give a single pitch estimate per frame by excluding the unvoiced and dual-pitch states from the dynamic programming stage.

Evaluation was restricted to voiced frames and a pitch estimate was classified as correct if it was within ±5% of the true value. Each of the graphs in Fig. 5 shows the performance of the algorithms for one of the noise types. It can be seen that at $+20$dB SNR, all of the algorithms reach

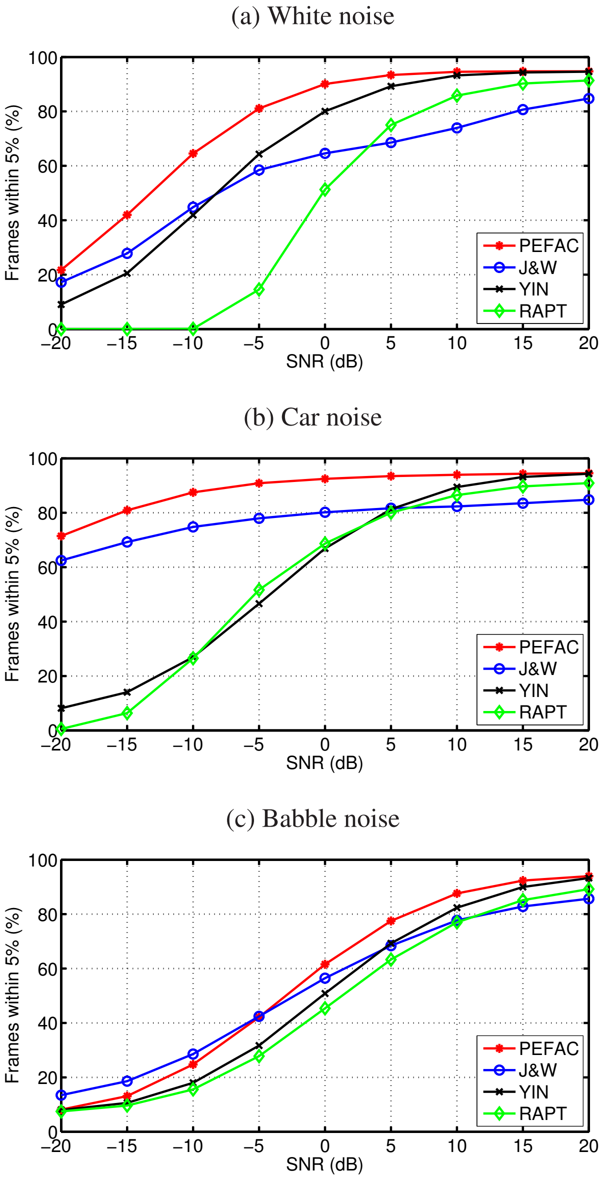


Figure 5: Variation of pitch estimation accuracy with SNR for (a) white noise, (b) car noise, and (c) babble noise. The solid lines show the percentage of correct frames (error below 5%) for each of the algorithms: PEFAC, J&W [15], YIN [14] and RAPT [13].

a performance plateau which varies between algorithms. The two time-domain algorithms, YIN and RAPT, degrade rapidly for all noise types at around 0 dB SNR although YIN always outperforms RAPT, particularly for white noise. The proposed algorithm (PEFAC) has excellent performance at $+20$ dB SNR and retains this high performance at significantly worse SNR levels than the other algorithms. The J&W algorithm degrades more gradually than the other algorithms and below -5 dB SNR it is the best algorithm for babble noise although at this level, all algorithms perform very poorly. Overall the performance of PEFAC consistently exceeds that of the other algorithms.

The RAPT and J&W algorithms employ dynamic programming to enforce soft temporal continuity constraints.

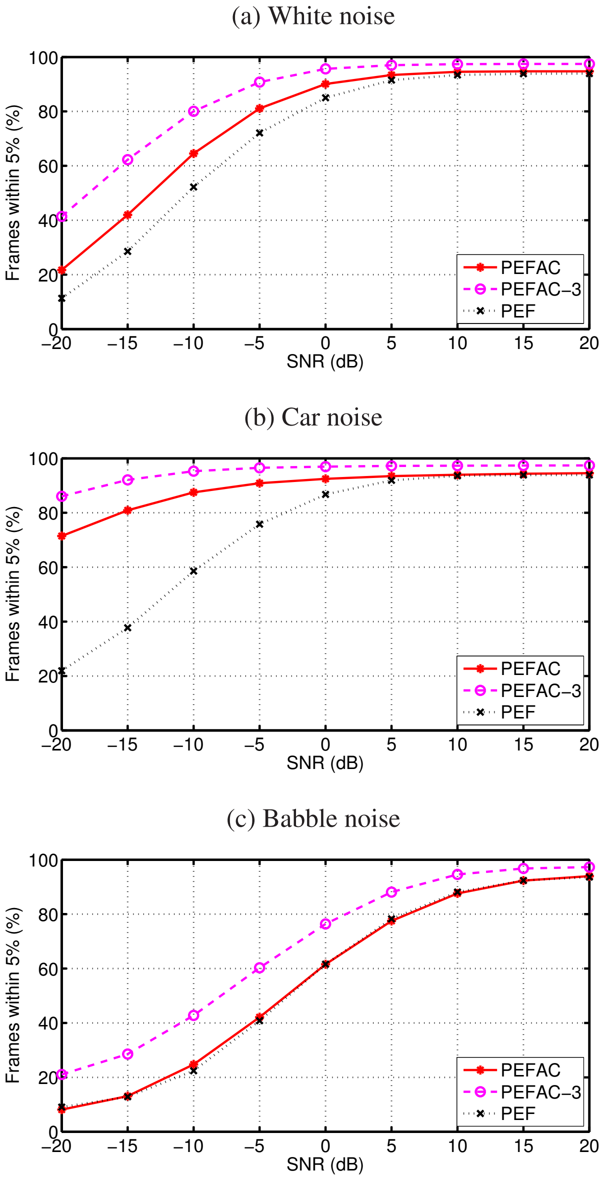


Figure 6: Variation of pitch estimation accuracy (error below 5%) with SNR for (a) white noise, (b) car noise, and (c) babble noise. The dashed line shows the frequency of the correct pitch being one of the top three PEFAC candidates (PEFAC-3). The solid line shows the percentage of correct frames for PEFAC. The dotted line shows the performance of the algorithm without amplitude compression (PEF).

Such constraints can be particularly effective at suppressing the octave errors that pitch estimators sometimes make. We have not used such constraints in this work but, as an indication of how they might improve performance, we have included as the dashed line in Fig. 6 (PEFAC-3) the percentage of frames for which the correct pitch was one of the three highest peaks in the filter output. From this we see that combining PEFAC with a perfect candidate selection algorithm could potentially give an additional performance improvement corresponding to 5 dB SNR. In Fig. 6 we can also observe the performance of the algorithm without the amplitude compression stage, represented with the dotted line (PEF). For white noise, below 0 dB SNR we get 10% improvement using amplitude compression. The improvement is even more visible for narroband noise such as car noise, going from 21.94% to 71.47% for -20dB SNR. The compression has no effect on babble noise, as the spectrum shape of the noise is similar to the speech spectrum.

5. CONCLUSIONS

In this paper we have presented the PEFAC pitch estimation algorithm and shown that it is able to give reliable pitch estimations even at poor SNRs. The algorithm comprises an amplitude compression stage that attenuates narrowband noise components with a pitch estimation filter that rejects broadband noise having a smooth power spectrum. The algorithm has been evaluated on the TIMIT core test set with a variety of noise types and consistently outperformed other widely used algorithms, even those that incorporate temporal continuity constraints.

REFERENCES

[1] D. Sharma and P. A. Naylor, "Evaluation of pitch estimation in noisy speech for application in non-intrusive speech quality assessment," in *Proc European Signal Processing Conf*, Aug. 2009, pp. 2514–2518.

[2] T. W. Parsons, "Separation of speech from interfering speech by means of harmonic selection," *J. Acoust Soc Amer*, vol. 60, no. 4, pp. 911–918, Oct. 1976.

[3] M. R. Schroeder, "Period Histogram and Product Spectrum: New Methods for Fundamental-Frequency Measurement," *J. Acoust Soc Amer*, vol. 43, no. 4, pp. 829–834, Apr. 1968.

[4] P. Martin, "Comparison of pitch detection by cepstrum and spectral comb analysis," in *Proc IEEE Intl Conf Acoustics, Speech and Signal Processing*, May 1982, vol. 7, pp. 180 – 183.

[5] D. J. Hermes, "Measurement of pitch by subharmonic summation," *J. Acoust Soc Amer*, vol. 83, no. 1, pp. 257–264, Jan. 1988.

[6] J. C. Brown, "Musical fundamental frequency tracking using a pattern recognition method," *J. Acoust Soc Amer*, vol. 92, no. 3, pp. 1394–1402, Sept. 1992.

[7] D. Byrne, H. Dillon, K. Tran, S. Arlinger, K. Wilbraham, R. Cox, B. Hagerman, R. Hetu, J. Kei, C. Lui, et al., "An international comparison of long-term average speech spectra," *J. Acoust Soc Amer*, vol. 96, no. 4, pp. 2108–2120, Oct. 1994.

[8] M. Brookes, "VOICEBOX: A speech processing toolbox for MATLAB," `http://www.ee.imperial.ac.uk/hp/staff/dmb/voicebox/voicebox.html`, 1997.

[9] J. S. Garofolo, "Getting started with the DARPA TIMIT CD-ROM: An acoustic phonetic continuous speech database," Tech. Rep., National Institute of Standards and Technology (NIST), Gaithersburg, Maryland, Dec. 1988.

[10] P. Boersma and D. Weenink, "Praat, a system for doing phonetics by computer (Version 5.1.23)," `http://www.fon.hum.uva.nl/praat/`, 2010.

[11] H. J. M. Steeneken and F. W. M. Geurtsen, "Description of the RSG-10 noise data-base," Tech. Rep. IZF 1988-3, TNO Institute for perception, 1988.

[12] ITU-T, "Objective measurement of active speech level," Recommendation P.56, Mar. 1993.

[13] D. Talkin, "A robust algorithm for pitch tracking (RAPT)," in *Speech Coding and Synthesis*, W. B. Kleijn and K. K. Paliwal, Eds., pp. 495–518. Elsevier, 1995.

[14] A. de Cheveigné and H. Kawahara, "YIN, a fundamental frequency estimator for speech and music," *J. Acoust Soc Amer*, vol. 111, no. 4, pp. 1917–1930, Apr. 2002.

[15] Z. Jin and D. L. Wang, "A multipich tracking algorithm for noisy and reverberant speech," in *Proc IEEE Intl Conf Acoustics, Speech and Signal Processing*, Mar. 2010, pp. 4218–4221.