Estimation of Speaker Individual Spectral Envelope for Pitch Tracking Improvement

Yaniv Zonis, Yaakov Buchris and Israel Cohen
Andrew and Erna Viterbi Faculty of Electrical Engineering
Technion - Israel Institute of Technology, Technion City, Haifa 32000, Israel
(szonis@tx , bucris@tx , icohen@ee).technion.ac.il

***Abstract*—Pitch estimation has been of great interest for several decades due to many important audio applications, such as music transcription, source separation, and speech coding. There are several approaches in the literature for estimating pitch, many of which make use of short-time spectrum analysis. A recently proposed algorithm, namely the PEFAC algorithm, performs pre-enhancement of speech components in the short time spectrum to yield a robust pitch estimation. This pre-enhancement procedure is based on a function that outlines the spectral envelope of human speech in the universal sense. In this paper, we propose to overcome some limitations of the PEFAC algorithm by employing an alternative enhancement procedure, which uses an estimation of the individual spectral envelope instead of using a universal function. This approach allows better correspondence to the specific speaker's spectral features. Experimental results show that the proposed algorithm outperforms the original PEFAC algorithm, especially in hard conditions such as low SNR and transient noise.**

I. Introduction

Fundamental frequency (pitch) estimation of audio signals has been of great interest for several decades, and plays an important role in applications such as music transcription, source separation, speaker recognition and speech coding. Pitch estimation algorithms are often implemented in the time-frequency domain, usually via short-time Fourier transform, by estimating the pitch for each frame individually. In an algorithm proposed by Schroeder [1] the spectrum of each frame is compressed in the frequency axis by several integer factors and all of the compressed versions are summed along with the original, so that harmonies of the concurrent fundamental frequency are 'piled up' on the first harmony peak, which then makes it easier to detect. In another method, proposed by Martin [2], each frame's spectrum is correlated with various comb filters of different base intervals and the pitch is chosen according to the maximally correlated comb filter. The method proposed by Hermes [3] takes advantage of the logarithmic scale's properties. Instead of compressing the frequency scale as done by Schroeder, the spectrum is transformed to the log-frequency scale and then merely shifted and summed to give the equivalent result of stacking harmonies on the fundamental frequency. Liu and Lin [4] proposed an algorithm where the pitch is determined for each frame's spectrum according to a score calculated for a range of candidate frequencies. The score function for each candidate frequency is derived from several measures of energy in prefixed vicinities of its harmonies.

Indeed many pitch tracking algorithms show impressive performance while utilizing the spectral behavior of harmonic signals, but there are hardly any attempts to incorporate assumptions derived from the signal being human speech, which might significantly improve the performance. Utilization of human speech characteristics was recently deployed in a pitch tracker called PEFAC [5]. The PEFAC algorithm, which also operates in the log-frequency domain, presents a novel pre-enhancement method (referred to as the *normalization stage*) which is based on the assumption that the spectral envelopes of any human's voiced syllables are close in shape to a universal and constant spectral envelope. The normalization stage in PEFAC is an essential contributor to its high performance, as supported by experiments. After the normalization, the spectrum is convolved with a matched filter designed to have a high correlation when aligned with the fundamental peak in the log-frequency domain, while having a much lower correlation when aligned with a higher harmony or a sub-harmony of the fundamental peak, and a much lower correlation when not aligned with any harmony. In addition to pitch estimation, the PEFAC algorithm provides an estimation for voiced-speech probability at each frame.

PEFAC outperforms some well-known pitch trackers, especially in high levels of noise, but when closely observing the effects of the normalization stage in PEFAC on speech frames, it is often found to adversely affect the coherence of the speech harmonies. In this paper, we propose to use PEFAC with an alternative normalization stage that makes a weaker assumption about the spectral envelope of the subject's voiced syllables. Instead of assuming a universal spectral envelope that supposedly resembles human speech in general, we estimate the subject speaker's spectral envelope and use it in place of the universal one. The proposed alternative normalization yields better results than the original PEFAC in the vast majority of tests performed in this work.

This Research was supported by Qualcomm Research Fund and MAFAAT - Israel Ministry of Defense.

978-1-5090-2152-9/16/$31.00

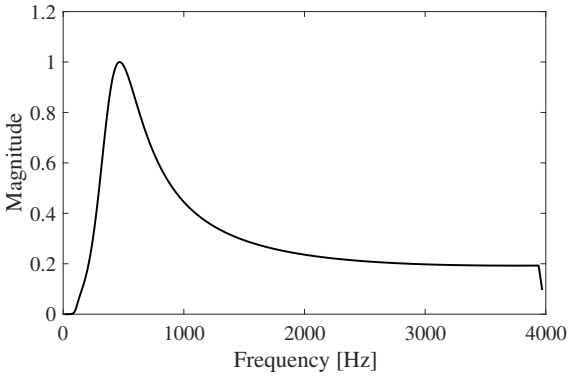

Fig. 1: The universal LTASS

II. SPECTRAL ENHANCEMENT

A. PEFAC's Normalization Algorithm

Byrne et al. [6] showed that averaging the short-time spectrum of a clean speech $S_t(f)$ along the time axis over a long period of time, produces a shape that strongly resembles a constant universal shape referred to in [6] as the Universal LTASS (Long Term Average Speech Spectrum). Accordingly, in frequencies where the LTASS has low power, it is less likely for a speech harmony to be present. Conversely, the more powerful frequency ranges of the LTASS are more likely to hold speech harmonies. Figure 1 depicts the universal LTASS used in PEFAC. Given a short-time spectrum of a noisy speech signal

$$Y_t(f) = S_t(f) + N_t(f), \tag{1}$$

where $S_t(f)$ is the clean speech and $N_t(f)$ is the noise, we can enhance it by multiplying it with the LTASS:

$$\widetilde{Y}_{t\mathrm{ENH}}(f) = Y_t(f)L(f). \tag{2}$$

This enhancement is expected to attenuate noises that appear in the low-power, low-likelihood frequency range of the LTASS. There are two major downsides to this approach. Firstly, in the presence of coloured noise whose power is mostly concentrated where the LTASS is low, transient noises might mount this noise and gain a head-start over speech harmonies. As a result, $\widetilde{Y}_{t\mathrm{ENH}}(f)$ will contain both levelled speech and transient components, instead of attenuating the transient below the speech power. Secondly, on occasion, there might be a transient noise component with such high power that would still keep it rivalling the speech harmonies after the enhancement in (2). To overcome these issues, PEFAC's normalization divides the short-time spectrum (1) by a smoothed version of itself prior to multiplying it with the LTASS. The smoothed spectrum is given by:

$$Y_{t\mathrm{SM}}(f) = Y_t(f) * H_{\mathrm{SM}}(t, f), \tag{3}$$

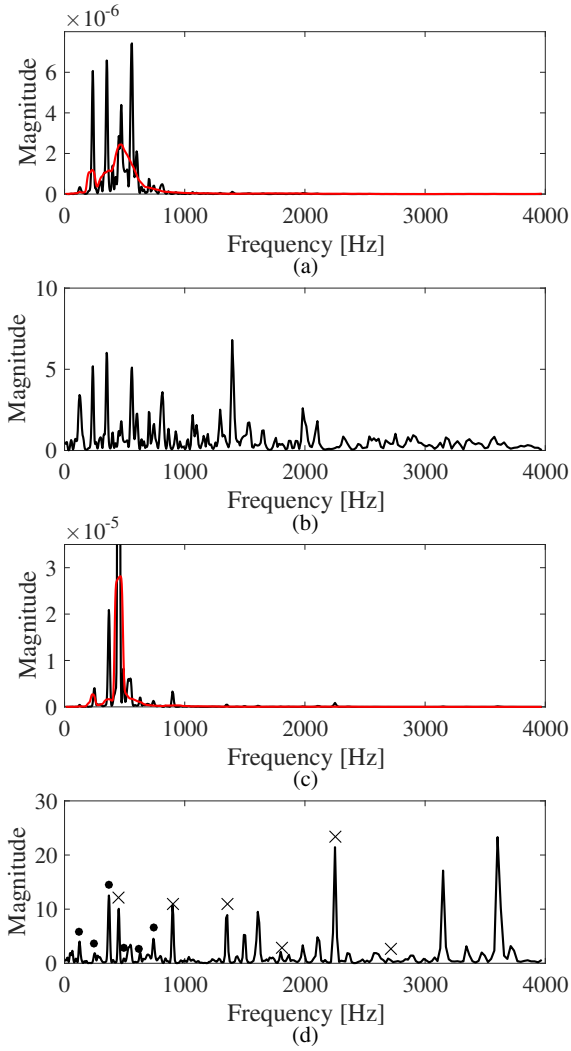

Fig. 2: The effect of division by smoothed spectrum. (a) Speech in presence of coloured noise - spectrum (black) and smoothed spectrum (red), (b) the normalized signal $Y_t(f)/Y_{t\mathrm{SM}}(f)$, (c) speech in presence of coloured noise and a powerful disturbing tone - spectrum (black) and smoothed spectrum (red), and (d) the normalized signal.

where '$*$' denotes a two dimensional convolution, and $H_{\mathrm{SM}}(t, f)$ is a uniform, rectangular filter. Thus, the normalization stage is as follows:

$$Y_{t\mathrm{ENH}}(f) = Y_t(f)\frac{L(f)}{Y_{t\mathrm{SM}}(f)}. \tag{4}$$

The smoothed spectrum acquires the shape of the stationary component of the background noise, thus dividing the spectrum by the smoothed spectrum cancels out the unbalancing effect that the stationary noise might have on the speech and transient components. Additionally, transient noises and harmonic speech components impose bumps on the smoothed spectrum, proportional in size to the transient or speech

magnitude. Consequently, transient noises and harmonic components are levelled amongst themselves after division by the smoothed spectrum.

Figures 2(a) and (b) show the effect of the division by the smoothed spectrum on a voiced frame in the presence of stationary, coloured noise. Figure 2(a) shows the original spectrum (black) along with the smoothed spectrum (red). The rise of the smoothed spectrum around 500 Hz shows its reaction to the stationary coloured noise, and the speech harmonies also impose a raised shape in the smoothed spectrum. Figure 2(b) shows the result of the division $Y_t(f)/Y_{t\mathrm{SM}}(f)$. Note that the scale of magnitude at Figure 2(b) is completely different than that of Figure 2(a) because of the division by the smoothed spectrum. The division caused the energetic harmonies seen in Figure 2(a) to level among themselves, while levelling several more speech harmonies that were not visible beforehand.

Figures 2(c) and (d) show the division effect on a voiced speech frame in the presence of the same coloured noise with the addition of a powerful harmonic tone with a base frequency of 450 Hz. Figure 2(c) shows the original spectrum (black) along with the smoothed spectrum (red). In this graph, most of the speech harmonies are hardly visible, due to the scale of the powerful tone at 450 Hz. A significant reaction to the powerful tone is seen in the smoothed spectrum, and a more moderate reaction to the speech harmonies. In Figure 2(d) we see the result of the division. The division is evidently not as effective in levelling the spectrum as it is in Figure 2(b), although the speech harmonies and the disturbing tone, marked here with dots and x's respectively, are now in the same scale, which is a much preferable situation than before the division.

Figure 3 shows the effect of multiplying the divided spectrum by the LTASS. The subjects displayed through Figures 3(a) to (d) correspond to those in Figures 2(a) to (d) respectively. In Figure 3(a) the LTASS seems to miss the first harmonies of speech, and instead drastically attenuates the first harmony as seen in Figure 3(b). Looking at the results of the product with the LTASS in Figures 3(c) and (d), it seems to have been beneficial in terms of attenuating the disturbing harmonic tone, marked with 'x', but like with the subject in Figures 3(a) and (b), the lowest of the speech harmonies, marked with dots, suffered an attenuation after product with the LTASS. These side effects motivate us to use a different spectral envelope and establish its individual estimation to be presented in the next section.

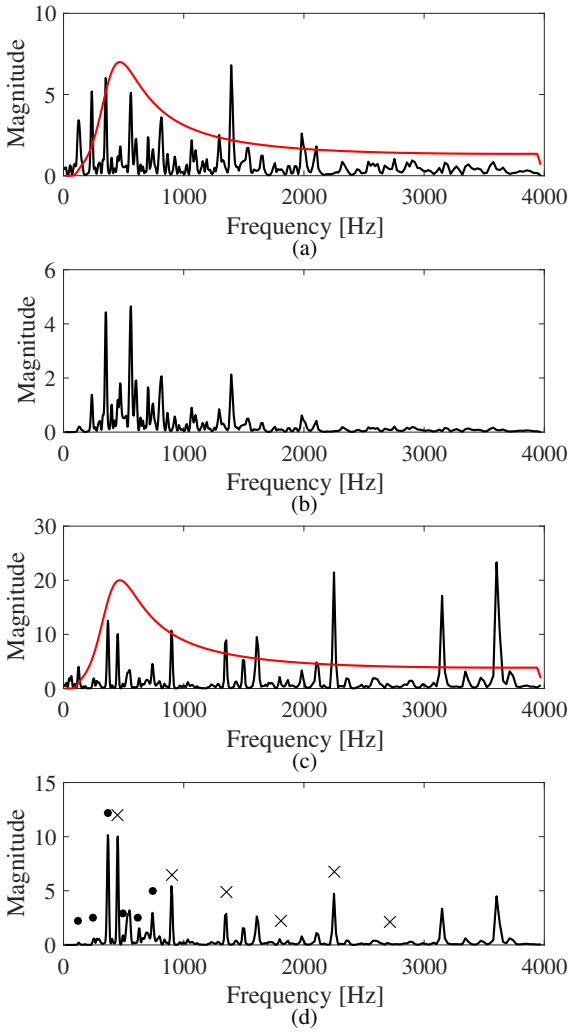

Fig. 3: The effect of product by LTASS. (a) Speech in presence of coloured noise - smoothed spectrum (black) and universal LTASS (red), (b) the product result $Y_{t\mathrm{ENH}}(f)$, (c) speech in presence of coloured noise and a powerful disturbing tone - smoothed spectrum (black) and universal LTASS (red), and (d) the product result.

B. Individual Envelope Estimation

The universal LTASS is not always beneficial as a normalization target function. In Figure 3, for example, the universal LTASS attenuated desirable speech harmonies and did not attenuate enough the undesirable spectral components. This is due to the assumption that the universal LTASS would adequately apply for all speakers, which imposes an adverse effect on recorded speakers whose individual LTASS significantly varies from the universal one. It is therefore preferable to estimate the subject speaker's individual spectral envelope and use it in place of the universal LTASS. Suppose we have $Y_{t\mathrm{IND}}(f)$ - a recording of the subject speaker in silence. The speaker's individual LTASS can be obtained by:

$$L_{\mathrm{I}}(f) = \frac{1}{N}\sum_{t} Y_{t\mathrm{IND}}(f), \tag{5}$$

where N is the number of frames in the short-time spectrum. In order to achieve a spectral envelope estimation that will better serve its purpose as an enhancing agent, some operations are applied on each speech frame prior to averaging through time. First, in order to prevent speech segments where the speaker speaks quietly or loudly to have an unequal effect on the estimated spectral shape, each speech frame is divided by the square root of its speech band energy, where the speech band energy is calculated using the universal LTASS as a weighting window:

$$\overline{Y}_{t\mathrm{IND}}(f) = \frac{Y_{t\mathrm{IND}}(f)}{\sqrt{\sum_f Y_{t\mathrm{IND}}^2(f) L(f)}}. \tag{6}$$

Then, using the original PEFAC algorithm for pitch estimation, each frame is masked to leave only the concurrent harmonies and attenuate undesirable regions of the spectrum. A most appropriate masking function is the one used in the PEFAC article as the matched filter used for finding the pitch, as described in Section I. As mentioned before, the PEFAC algorithm operates in the log-frequency domain, which is helpful in our case also because it allows the use of the same matched filter at every frame, regardless of the concurrent pitch. The matched filter from [5] is defined by:

$$h(q) = \frac{1}{\gamma - \cos\left(2\pi e^q\right)} - \beta, \tag{7}$$

for $\log(0.5) < q < \log(K + 0.5)$ and $h(q) = 0$ otherwise, where q is the log-frequency, γ controls the peaks width, K is the number of harmonies expected to be present, and β is chosen so that $\int h(q)dq = 0$. Finally, we would like to weight the frames according to the probability of them containing voiced speech. For this purpose, we make use of the PEFAC algorithm's output of voice probability estimation. Mathematically, the spectral envelope estimation is given by:

$$L_{\mathrm{I}}(q) = \frac{1}{\sum_t \tilde{v}_t} \sum_t \overline{Y}_{t\mathrm{IND}}(q) h(q - \log(\tilde{f}_t))\tilde{v}_t, \tag{8}$$

where $\tilde{v}_t \in [0, 1]$ is the voice probability estimation at time frame t, and $\tilde{f}_t$ is the pitch estimation at time frame t. This estimated spectral envelope can now benefit the PEFAC algorithm when processing recordings of the subject speaker, with or without the presence of noise.

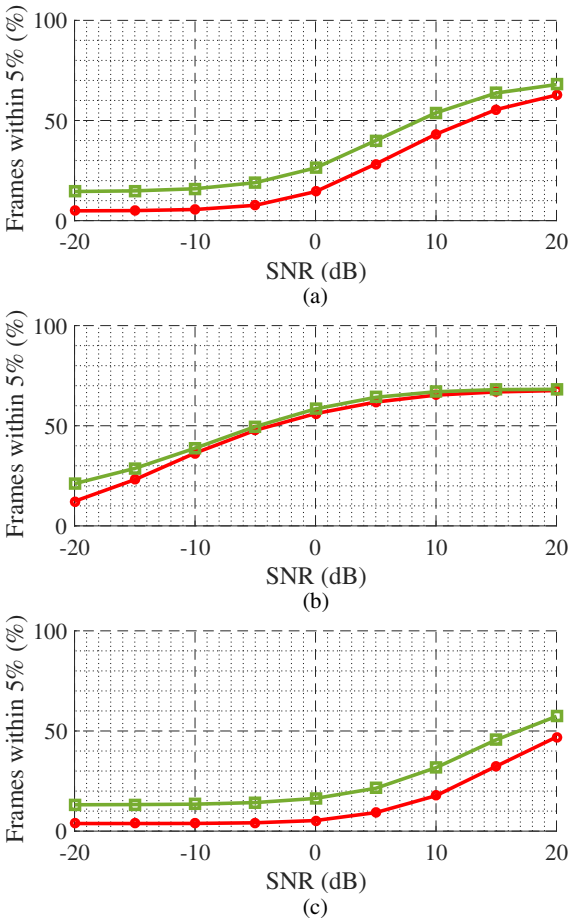

Fig. 4: PEFAC's performance with the original normalization (red) and the alternative normalization (green) in different noise conditions. (a) White noise, (b) car noise, and (c) babble noise.

III. Experimental Results

The alternative normalization algorithm was evaluated using PEFAC [5], [7]. Speech samples and respective ground-truth pitch values were taken from Bagsaw's database [8]. This database contains 50 samples of one male speaker and 50 samples of one female speaker. Each speaker has a total of about 60 seconds of active speech. Noises from the NoiseX92 database [9] were used, namely car (vehicle interior), babble and white. Different SNR conditions were created when speech level was determined using ITU-T P.56 [10], [7] and noise level was determined using unweighted power. Universal LTASS values were taken from Table II in [6].

For each speaker, 35 sentences were used for estimating the spectral envelope while using the clean version, and the remaining 15 sentences were used for evaluating the proposed algorithm while imposing different noise conditions. For better statistics, 10 different partitions were applied when the first uses sentence 1 through 35 for estimation, the second uses sentences 6 through 40 for estimation, the third 11 through 45, and thus further cyclically. The measurement used for

evaluation was the percentage of frames on which the pitch detector deviated no more than 5 percent from the ground truth value.

In Figure 4 we present the scores for each noise type and noise level, where each score is an average on all sentences and partition-combinations. Figure 4(a) shows the results on white noise, 4(b) shows the results on car noise, and 4(c) on babble noise. Figure 5 shows the percentage of runs on which the estimated spectral envelope gave better results than the universal LTASS. The results show a clear advantage for using an individually estimated spectral envelope over the universal LTASS in the presence of babble and white noise, though car noise seems to impose a greater difficulty which leaves the proposed alternative at a significant advantage only in the lowest SNR conditions.

IV. Conclusion

We have proposed an improvement to the spectral enhancement process that was integrated into the PEFAC algorithm. The proposed method requires estimation of the subject speaker's individual spectral envelope. Assuming a recording of the subject's clean speech is available, we have introduced an estimation process which is based on the LTASS mathematical definition. The proposed alternative spectral enhancement achieved superior results compared to those of the original PEFAC through most conducted experiments, except the experiments with car noise which are not conclusive. An interesting option for further research is weighting the estimated spectral envelope with the universal LTASS to overcome the weaknesses of either method. Also, the estimation of the speaker's spectral envelope in noise conditions should be further investigated.

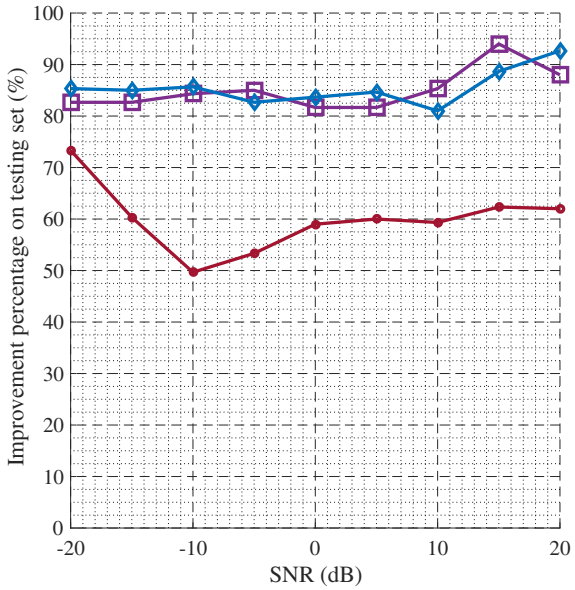

Fig. 5: The improvement rate of the proposed spectral envelope estimation over the universal LTASS, in presence of different noise types. White noise (purple line with squares), car noise (red line with dots) and babble noise (blue line with diamonds).

References

[1] M. R. Schroeder, "Period histogram and product spectrum: New methods for fundamental-frequency measurement," *The Journal of the Acoustical Society of America*, vol. 43, no. 4, pp. 829–834, 1968.

[2] P. Martin, "Comparison of pitch detection by cepstrum and spectral comb analysis," *in Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 180–183, 1982.

[3] D. J. Hermes, "Measurement of pitch by subharmonic summation," *The journal of the Acoustical Society of America*, vol. 83, no. 1, pp. 257–264, 1988.

[4] D.-J. Liu and C.-T. Lin, "Fundamental frequency estimation based on the joint time-frequency analysis of harmonic spectral structure," *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 6, pp. 609–621, 2001.

[5] S. Gonzalez and M. Brookes, "PEFAC-a pitch estimation algorithm robust to high levels of noise," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 2, pp. 518–530, 2014.

[6] 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," *The Journal of the Acoustical Society of America*, vol. 96, no. 4, pp. 2108–2120, 1994.

[7] M. Brookes *et al.*, "VOICEBOX: Speech processing toolbox for MATLAB," *Software, available [Mar. 2011] from www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html*, 1997.

[8] P. Bagshaw, "Paul bagshaws database for evaluating pitch determination algorithms," *www.cstr.ed.ac.uk/research/projects/fda*, vol. 499, p. 500.

[9] A. Varga and H. J. Steeneken, "Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech Communication*, vol. 12, no. 3, pp. 247–251, 1993.

[10] I. Rec, "P. 56, objective measurement of active speech level," *International Telecommunication Union, CH-Geneva*, 1993.