A Pitch-Based Spectral Enhancement Technique for Robust Speech Processing

Kantapon Kaewtip [1], *Lee Ngee Tan*[1], *Abeer Alwan*[1]

[1]Department of Electrical Engineering, University of California, Los Angeles, California, USA

kkaewtip@ucla.edu, tleengee@ee.ucla.edu, alwan@ee.ucla.edu

Abstract

This paper presents a new pitch-based spectral enhancement algorithm on voiced frames for speech analysis and noise-robust speech processing. The proposed algorithm determines a time-warping function (TWF) and the speaker's pitch with high precision, simultaneously. This technique reduces the smearing effect in between harmonics when the fundamental frequency is not constant within the analysis window. To do so, we propose a metric called the *harmonic residual* which measures the difference between the actual spectrum and the resynthesized spectrum derived from the linear model of speech production with various combinations of TWF and high-precision pitch values as parameters. The TWF and pitch pair that yields the minimum harmonic residual is selected and the enhanced spectrum is obtained accordingly. We show how this new representation can be used for automatic speech recognition by proposing a robust spectral representation derived from harmonic amplitude interpolation[1].

Index Terms: speech recognition, feature extraction, noise robust, pitch

1. Introduction

With the advent of fast processor chips, automatic speech recognition (ASR) applications are more widely used today compared to a decade ago [1]. ASR works well in quiet environments, but a large degradation in performance is observed when the speech signal is noisy [2]. This is because the noise distorts the spectral shape of the speech spectra, from which cepstral features (e.g. Mel-frequency cepstral coefficients - MFCCs) for ASR are extracted. [3]

To reduce such distortions, various spectral enhancement techniques have been proposed to improve the noise robustness of ASR [3-8]. For example, PNCC features [9], attempt to find a noise floor based on the energy profile and then performs normalized power bias subtraction. LSEN [10] performs spectrogram enhancement by multiplying the noisy spectrogram with a smoothed SNR-weighted mask. These techniques suppress the low SNR regions which are prone to noise corruption, and are highly dependent on the accuracy of the noise estimation technique used for SNR computation [11]. It is well-known that spectral harmonics in voiced speech frames have high energies and thus are more robust to noise than other parts of the spectrum [12-14]. Many noise robust speech applications based on spectral harmonic components have been developed [15-18]

In non-tonal languages, such as English and most European languages, it is believed that most information for ASR is contained in the vocal tract transfer function, i.e. the envelope of the spectrum. Several attempts were made to capture the vocal tract transfer function by extracting the envelope of the spectrum such as LPC-based techniques [19-21] and [22]. However, without any noise compensation, the envelope usually contains spurious peaks when the noise level is high. By using the fundamental frequency (f0), however, the noise harmonics can be distinguished so that only the speech harmonics are used to compute the spectral envelope.

Harmonic frequency can, for example, be found by sequential detection based on the initial f0 estimate [15,23] such that $f_k = f_{k-1} + f0 + f_k$, where f_k is the k-th harmonic and f_k is a small frequency adjustment added to select the highest spectral magnitude component in the neighborhood of $f_{k-1} + f0$. The problem of this method is that harmonic tracking can be difficult in low SNR cases, making it challenging to identify subsequent harmonics at high frequencies. Another way to identify harmonic frequencies is to assume that they fall in the neighborhood of kf0, i.e. $f_k = kf0 + f_k$, which requires re-estimation of f0 to higher precision in order to avoid erroneous estimation of harmonic locations in the high frequency region. To illustrate this problem, suppose the actual f0 is 100 Hz but the estimated f0 is 102. By the 25th harmonic, the frequency will be off by 50Hz so the identified harmonic frequency is at exactly the middle of harmonic 25th and 26th. If the search range of the peaks from kf0 is more relaxed, a noise peak might be confused with the harmonic. Therefore, a noise-robust harmonic localization algorithm is needed to re-estimate f0 effectively.

Harmonic peak localization, however, is also difficult when the harmonic structure is "smeared" in the presence of a rapidly-changing f0 such that the harmonic structure in the high frequency region is corrupted [24]. In this case, the inter-peak samples may have high amplitudes, and some false peaks might be generated due to aliasing and superposition of harmonics. In this case, clear harmonic structure can be obtained by warping the signal to increase the periodic structure, as can be done with the Chirp transform (ChT) [25-28]. Applications of the ChT, include speech analysis, high-resolution pitch detection [29] and music representation[30-31] as the ChT can enhance the harmonic structure. However, ChT has not been used in pitch-based noise estimation and ASR due to the computational complexity of the transform. Moreover, for a noisy signal, an accurate chirp rate is difficult to estimate.

In this paper, we present a noise robust technique that determines the optimal TWF and harmonic locations simultaneously (Section 2). We also show low-complexity implementations of the time warping procedure and harmonic localization algorithm and discuss how it can be used to improve noise robust ASR features. In Section 3, we briefly describe our experimental setup on the Aurora 2 database. Section 4 presents the results with a discussion, while Section 5 provides a summary and future work

[1] The work is supported in part by DARPA

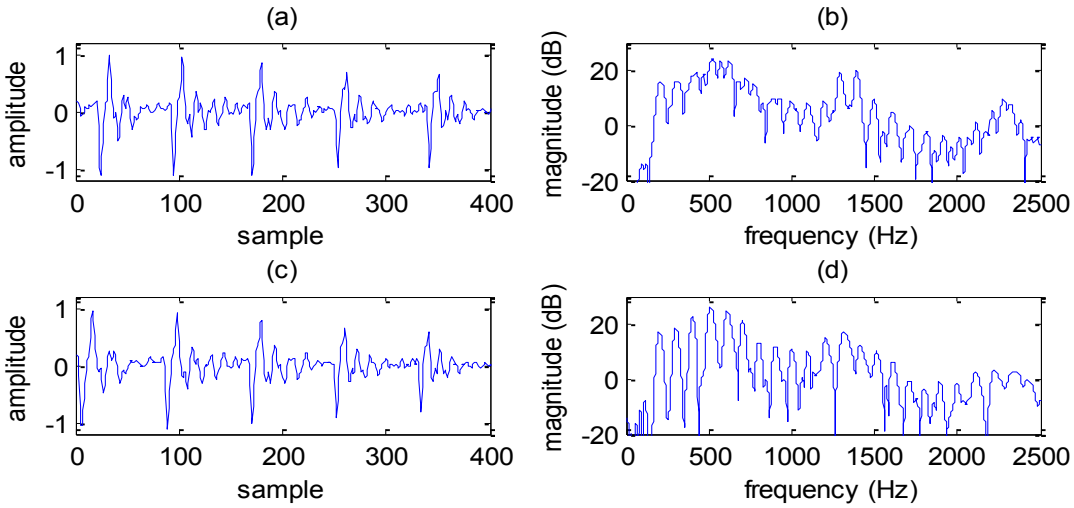

Figure 1. Illustration of a time-warping function. Part (a) and (b) are the time waveform and frequency spectrum of the original signal, respectively. Part (c) and (d) are the same as (a) and (b) but after time warping.

2. Proposed Algorithm

To obtain a refined harmonic spectrum, the TWF and high-precision f0 need to be estimated. These two parameters are solved simultaneously using a metric called *a harmonic residual* which measures the difference between the refined actual spectrogram and harmonic spectral representation obtained via TWF.

2.1 Time-Warping the Time-Domain Signal

The algorithm has a set of time warping functions (TWF) each of which maps values from the signal x(n) to the output signal $x_\alpha(n)$ by the relationship

$$x_\alpha(n) = x(n+\alpha n^2) \qquad (1)$$

where the reference n = 0 is the middle sample. For example, with $\alpha = 0.0002$, $x_\alpha(100) = x(100+2)$ meaning the sample n=100 is moved to the right by 2 samples. Since the TWF is derived from a continuous time function, if the argument of x_α is a non-integer, the value is approximated by linear interpolation. For example, $x_\alpha(102.4)$ is approximated by $x_\alpha(102) + 0.4(x_\alpha(103)-x_\alpha(102))$. We found that the range of α used for human pitch profile within the range of -0.0008 to 0.0008. This range maybe adjusted to other applications such as music representation [31].

Figures 1.a and 1.b represent x(n) and its FT magnitude (with a Hamming window), respectively. This clean signal has a varying period as shown in Figure 1.a but its harmonic structure is distorted (Figure 1.b). Now, let us give an example of how time warping can enhance the harmonic representation. When a TWF with $\alpha = 0.00048$ is applied to the signal, the output after warping is shown in 1.c and the corresponding spectrum is shown in 1.d. It can be seen that each period in 1.c. becomes equal and the harmonic structure in 1.d is enhanced. In order to find the optimal TWF or α, we propose a metric to measure the harmonicity of the warped output signal as described in Section 2.2.

2.2 Resynthesizing the refined harmonic spectrum

In STFT, when a signal ***s*** is multiplied by a window ***w***, the signal **S** convolves with the window **W** in the frequency domain with **W** normalized so that W(0) = 1. Let $\mathbf{a_k}$ be the amplitude at the k-th harmonic frequency, then the spectrum of a voiced frame (according to the linear modeling of speech production) can be described by

$$R(f) = \left(\sum_{k=1}^{p} a_k \delta(f - kf_0)\right) * W(f - kf_0) \qquad (2)$$

where f denotes the frequency index, p is the number of harmonics in the range of the signal bandwidth and (f) is an impulse function. By neglecting the phase for simplicity, the magnitude spectrum can be synthesized by

$$R(f) = \sum_{k=1}^{p} a_k |W(f - kf_0)| \qquad (3)$$

Let **T** be a Toeplitz matrix whose first column is the absolute value of the Fourier Transform of the normalized window W. The resynthesized spectral magnitude R(f) can be obtained by a single linear equation

$$R = T_a \mathbf{a} \qquad (4)$$

where **a** is the column vector of $\mathbf{[a_1\ a_2\ \ldots\ a_p]'}$ and $\mathbf{T_a}$ is a subsampled matrix whose columns correspond to the DFT index closet to the of harmonic frequencies. Intuitively, the equation says that the amplitude at a frequency f is a linear combination of all harmonic amplitudes with weights corresponding to the distance to that frequency. If the f0 candidate value is accurate, at high SNRs, the harmonic spectrum resynthesized using the optimum TWF would be very similar to the original DFT spectrum. We call the L_1 norm of the difference between the original and the resynthesized spectrum, the *harmonic residual,* which is defined by

$$rsd(f_i,\alpha) = \|S_\alpha - R_{\alpha,fi}\|_{L1} \qquad (5)$$

where f_i is the i-th f0 candidate, α is the parameter in Equation (1), $\mathbf{S_\alpha}$ is the original spectrum of $x_\alpha(n)$, and $\mathbf{R_{\alpha,fi}}$ is the resynthesized spectrum using f_i as the pitch hypothesis.

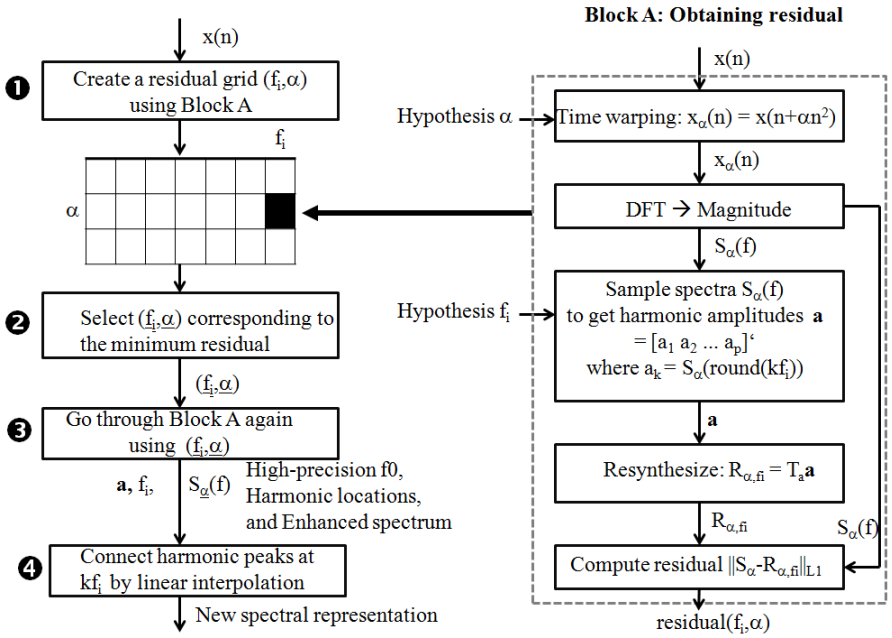

Figure 2. The overview of the proposed algorithm. Signal x(n) is a short-time waveform at a given voiced frame. Block A describes how a residual as a function of α and f_i is computed. The block diagram on the left describes how the optimal $\underline{\alpha}$ and $\underline{f_i}$ are obtained and therefore the high-precision f0, enhanced spectrum and harmonic locations.

For a low-pitched frame whose harmonics are close to each other and there is high sidelobe aliasing, the residual will still be low making this method virtually independent of the harmonic spacing. In the presence of noise, the residual will be higher depending on the noise level. Since the components at the harmonic frequencies generally have high SNRs, we only compute the residual in the frequency range of 40 Hz around each harmonic frequency, so that the residual is less dependent on the noise profile. For each frame, the combination of the TWF and f0 that give the minimum residual are considered optimal, and consequently used to generate the refined harmonic spectral representation for noise robust processing.

Note that this resynthesis method is an approximation of the clean spectrum due to the assumptions we made. For example, the values a_i taken directly from the original spectrum might be influenced by noise components. However, we found that in most cases, this approximation yields a reasonable measurement of harmonicity.

2.3 Deriving an Enhanced Spectral Envelope Representation

To reduce the f0 search range, we use f0 estimate as an input from the noisy signal and use range from -0.0008 to 0.0008 with a precision of 0.00002 (a more efficient search grid can be obtained by limiting α range due to the f0 derivative). The noise-robust multi-band summary correlogram-based pitch detector described in [32] is used, with modifications to reduce computational complexity. A fixed window length of 60ms is used, and frame-by-frame mean subtraction from the pre-comb-filtered subband envelopes is omitted. The hypothesis of f0 is limited to the 20% range of the initial estimated f0 from the pitch tracker. Then linear interpolation connects each amplitude peak (a_1, a_2, ..., a_p)at harmonic frequencies of the refined harmonic spectrum obtained by TWF to derive a new spectrographic representation as illustrated in Figure 3. Once all the voiced frames are processed, the final spectrogram is then used for feature extraction. For unvoiced frames, for simplicity, no spectral enhancement is made.

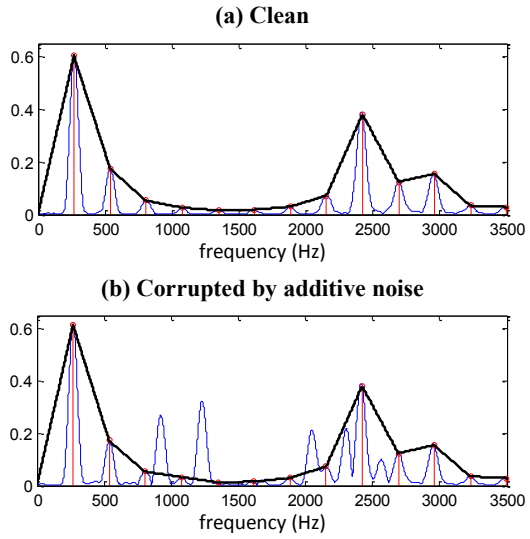

Figure 3. Linear interpolation of amplitudes at harmonic frequencies of refined spectrums. In Figure a) and b), the vertical lines represent harmonic amplitudes. The thick lines are linear envelopes obtained by our algorithm. Note that this procedure reduces the mismatch between two spectra taken from clean and noisy utterances.

3. Experimental Setup

We study the noise-robustness of the spectral representation by integrating it with some existing feature extraction methods, namely MFCC [33] and PNCC [9]. For MFCC, the enhanced spectrogram is converted into Mel filter bank

	Feature	Word Accuracy (%)					
		20dB	**15dB**	**10dB**	**5dB**	**0dB**	**Avrg**
1.	MFCC	97.6	93.6	78.7	54.8	11.9	65.5
2.	MFCC + linear envelope using initial estimated f0	95.0	86.3	64.6	39.8	20.2	61.2
3.	MFCC + time warping + linear envelope	97.6	95.4	88.8	72.7	43.8	79.6
4.	PNCC	98.7	97.3	93.3	81.1	53.7	84.8
5.	PNCC + linear envelope using initial estimated f0	96.5	94.4	88.9	76.0	53.5	81.9
6.	PNCC + time warping + linear envelope	98.2	96.6	92.5	82.1	60.8	86.0
7.	**PNCC + time warping + smooth envelope + spectral flooring**	**98.2**	**96.9**	**93.5**	**83.5**	**62.8**	**87.0**
8.	ETSI	98.1	96.7	92.8	83.2	59.8	86.1
9.	LSEN	98.3	97.1	93.9	83.2	59.6	86.4

Table 1. Word-accuracies for different front-ends on Aurora 2.

representation using 26 filters, followed by logarithmic compression and cepstral coefficients computation. The first 13 coefficients are retained and the first and second derivatives are computed. For PNCC, we found that the 2010 version with Power Bias Subtraction (PBS) Algorithm [9] yields better performance than the latest version of PNCC [34] on Aurora 2 dataset. For this reason, we employed PNCC with PBS in our study as follows. The enhanced spectrogram is passed to a gammatone filterbank filter of 40 channels. Next, power bias subtraction and normalization are applied and the enhanced filterbank values is scaled by power 1/15, followed by the same cepstral feature extraction as MFCC. We also compare the integrated features with two state-of-the-art algorithms namely LSEN [8] and ETSI [36].

Experiments have been conducted using the HTK-based back-end with the Aurora-2 corpus of noisy digits [35]. We model 11 words with 18-state 3-mix. HMMs. Two silence models are used with, respectively, 5 and 3 states, and 3 and 6 mixtures per state. Testing sets A and B comprise the same utterances corrupted by 8 types of background noise at various SNR from 0 dB to 20 dB.

4. Results and Discussion

Table 1 presents the performance in word accuracy percentage. The overall results can be summarized as follows. By simply taking spectral envelopes (using only the initial value of f0) before computing MFCC and PNCC without refining the spectrogram and f0 (Features 2 and 5), the accuracy drops, especially at high SNRs. Using time warping and estimating the spectral envelope as was described in the proposed algorithm (Features 3 and 6), the performance improves dramatically (compare with Features 1
and 4) especially at low SNRs. This is because the envelope obtained by interpolating the amplitude at the harmonics reduces the mismatch between training and testing samples because the new spectral representation is less dependent on the noise energy in between harmonics.

In this paper, we focus on the effect of simply interpolating the peaks at the harmonic frequencies from high precision f0 and enhanced spectra, but the performance can be improved further by making the new spectrographic representation compatible with the original features. In our preliminary experiment, we used a smooth envelope (as opposed to an envelope from linear interpolation). The smooth envelope is obtained by the resampling technique used in DSP. Since the harmonic frequencies are equally spaced by a value of f0, they can be viewed as decimated version of the spectral envelope. Therefore a sequence containing harmonic amplitudes is upsampled to get a low frequency envelope. The new envelope improves to 86.8% from the simple linear interpolation. We also found that by simple spectral flooring, the accuracy improves to 87%. The optimal configuration will be explored further in future work.

Note that our TWF method aims to find a better spectral representation in a similar way as ChT. Although our implementation is simpler, it has some drawbacks. For example, the linear interpolation of x(102.4) makes the algorithm easy to implement but has poor resolution in the high-frequency region. The set of TWFs is derived based on an assumption that the rate of f0 change is a linear function of time. However, we found that our implementation is sufficient in most cases. A more accurate representation can be obtained using a variety of the Chirp transforms and the Constant-Q Transform [31,37].

We have shown how the enhanced spectrum and high-precision f0 can be used to improve noise robust ASR features. However, the harmonic residual and the resynthesized-spectrogram, as by-products, also have potential applications in noise robust speech processing. The harmonic residual can be used to estimate noise in a similar way as the harmonic tunneling technique [38]. This noise estimation at voiced frames can extrapolate to unvoiced frame. Harmonic tunneling tends to over-estimate noise at high SNR due to window aliasing and the smearing effect but the harmonic residual in the proposed algorithm can reduce both undesirable effects. In addition, the resynthesized spectrum can be used for speech enhancement as noise in-between harmonics is discarded, reducing the musical perception which is commonly found in spectral subtraction enhancement. We will explore these two potential applications in the future.

5. Conclusions

In this paper, we propose an algorithm that enhances voiced spectra. For each voiced frame, a combination of a time warping function (TWF) and high-precision f0 values are considered. A resynthesized spectrum is obtained by a linear product of the harmonic amplitudes and the matrix generated by the DFT window. A residual that measures the difference between the actual and resynthesized spectrum is then computed. The combination of TWF and f0 that yields the minimum residual is selected. The optimal TWF is used to warp the signal so that the harmonic structure is less distorted and smearing effects are reduced. The optimal f0 is then used to locate harmonic frequencies. Preliminary experiments show that the proposed algorithm can be used to improve ASR performance in noise.

6. References

[1] Li Deng and Xuedong Huang. 2004. Challenges in adopting speech recognition. Commun. ACM 47, 1 (January 2004), 69-75.

[2] Y. Gong, "Speech recognition in noisy environments: A survey." Speech communication 16, no. 3 (1995): 261-291.

[3] Openshaw, J. P., and J. S. Masan. "On the limitations of cepstral features in noise." In Acoustics, Speech, and Signal Processing, 1994.

ICASSP-94., 1994 IEEE International Conference on, vol. 2, pp. II-49. IEEE, 1994.

[4] Sorensen, H.B.D. “Noise-robust speech recognition using a cepstral noise reduction neural network architecture Neural Networks”, IJCNN-91-Seattle, vol. 2, pp. 795–800, 1991

[5] Neumeyer, L.; Weintraub, M. “Robust speech recognition in noise using adaption and mapping techniques”, ICASSP-95,vol. 1, pp. 9–12, 1995 2

[6] Boll, S.F. “Supression of Acoustic Noise in Speech Using Spectral Subtraction”, IEEE Trans. Acoust., Speech, Signal Processing, vol. ASSP-27, pp. 113– 120, 1979 2

[7] Ephraim, Y; Malah, D “Speech Enhancement Using Minimum Mean Square Error Short-Time Spectral Amplitude Estimator”, IEEE Trans. On Acoustics, Speech, and Signal Processing, Vol. 32, No. 6, pp. 1109–1121, 1984 3

[8] B. Raj, M. L. Seltzer, and Richard M. Stern, “Reconstruction of missing features for robust speech recognition”, Speech Communication, vol. 43, pp. 275–296, 2004

[9] C. Kim and R.M. Stern, "Feature extraction for robust speech recognition based on maximizing the sharpness of the power distribution and on power flooring", ICASSP 2010, pp. 4574-4577.

[10] J.Hout and A. Alwan, "A Novel Approach to Soft-Mask Estimation and Log-Spectral Enhancement For Robust Speech Recognition", ICASSP 2012, pp. 4105-4108.

[11] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics", IEEE TASP, vol. 9, pp.504 - 512, 2001.

[12] Beh, Jounghoon, and Hanseok Ko. "A novel spectral subtraction scheme for robust speech recognition: spectral subtraction using spectral harmonics of speech." In Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03). 2003 IEEE International Conference on, vol. 1, pp. I-648. IEEE, 2003.

[13] Seltzer, Michael L., Jasha Droppo, and Alex Acero. "A harmonicmodel-based front end for robust speech recognition." In Proc. Eurospeech, vol. 3, pp. 1277-1280. 2003

[14] Wang, DeLiang, and Guy J. Brown, eds. Computational auditory scene analysis: Principles, algorithms, and applications. IEEE Press, 2006.

[15] Ealey, Douglas, Holly Kelleher, and David Pearce. "Harmonic tunnelling: tracking non-stationary noises during speech." In Proc. Eurospeech, vol. 1, pp. 437-450. 2001.

[16] Morales-Cordovilla, Juan A., Ning Ma, Victoria Sánchez, Jose L. Carmona, Antonio M. Peinado, and Jon Barker. "A pitch based noise estimation technique for robust speech recognition with missing data." In Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on, pp. 4808-4811. IEEE, 2011.

[17] Morales-Cordovilla, Juan A., Antonio M. Peinado, and Victoria Sánchez. "EQUIVALENCES AND LIMITS OF PITCH-BASED TECHNIQUES FOR ROBUST SPEECH RECOGNITION."

[18] Buera, Luis, Jasha Droppo, and Alex Acero. "Speech enhancement using a pitch predictive model." In Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on, pp. 4885-4888. IEEE, 2008.

[19] Hermansky, Hynek. "Perceptual linear predictive (PLP) analysis of speech." The Journal of the Acoustical Society of America 87 (1990): 1738.

[20] Thomas, Samuel, Sriram Ganapathy, and Hynek Hermansky. "Recognition of reverberant speech using frequency domain linear prediction." Signal Processing Letters, IEEE 15 (2008): 681-684.

[21] Athineos, Marios, and Daniel PW Ellis. "Frequency-domain linear prediction for temporal features." In Automatic Speech Recognition and Understanding, 2003. ASRU'03. 2003 IEEE Workshop on, pp. 261-266. IEEE, 2003.

[22] Q. Zhu and A. Alwan, "Amplitude Demodulation of Speech Spectra and its Application to Noise Robust Speech Recognition,"6th International Conference on Spoken Language Processing, ICSLP 2000. Vol. 1, pp. 341-344

[23] L. N. Tan, B. J. Borgstrom and A. Alwan, "Voice Activity Detection using Harmonic Frequency Components in Likelihood Ratio Test," ICASSP 2010, pp. 4466-4469.

[24] Xia, Xiang-Gen. "Discrete chirp-Fourier transform and its application to chirp rate estimation." Signal Processing, IEEE Transactions on 48, no. 11 (2000): 3122-3133.

[25] S. Mann and S. Haykin, “The chirplet transform: physical considerations,” IEEE Transactions on Signal Processing, vol. 41, no. 11, pp. 2745–2761, 1991.

[26] L. Weruaga and M. Képesi, “The fan-chirp transform for nonstationary harmonic signals,” Signal Processing, vol. 87, no. 6, pp. 1504–1522, 2007.

[27] Rabiner, L., R. Schafer, and C. Rader. "The chirp z-transform algorithm." Audio and Electroacoustics, IEEE Transactions on 17, no. 2 (1969): 86-92.

[28] Ozaktas, Haldun M., Billur Barshan, David Mendlovic, and Levent Onural. "Convolution, filtering, and multiplexing in fractional Fourier domains and their relation to chirp and wavelet transforms." JOSA A 11, no. 2 (1994): 547-559.

[29] Képesi, Marián, and Luis Weruaga. "High-resolution noise-robust spectral-based pitch estimation." submitted to Eurospeech (2005).
[30] M. Bartkowiak, “Application of the fan-chirp transform to hybrid sinusoidal+noise modeling of polyphonic audio,” in 16th European Signal Processing Conference, 2008.

[31] Cancela, Pablo, Ernesto López, and Martín Rocamora. "Fan chirp transform for music representation." In Proc of the 13th Int Conference on Digital Audio Effects DAFx10 Graz Austria, pp. 1-8. 2010.

[32] Lee Ngee Tan and Abeer Alwan, "Noise-Robust F0 Estimation Using SNR-Weighted Summary Correlograms From Multi-Band Comb Filters," ICASSP 2011, pp. 4464-4467.

[33] Zheng, Fang, Guoliang Zhang, and Zhanjiang Song. "Comparison of different implementations of MFCC." Journal of Computer Science and Technology 16, no. 6 (2001): 582-589.

[34] Kim, Chanwoo, and Richard M. Stern. "Power-normalized cepstral coefficients (pncc) for robust speech recognition." In Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on, pp. 4101-4104. IEEE, 2012.

[35] D. Pearce and H. G. Hirsch, “The aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions”, Automatic Speech Recognition: Challenges For the New Millennium, ASR2000, 2000, pp. 181–188.

[36] Speech Processing, Transmission and Quality Aspects (STQ); Distributed Speech Recognition; Adv. Frontend Feature Extraction Algorithm; Compression Algorithms, ETSI ES 202 050, 2007.

[37] J. C. Brown, “Calculation of a constant Q spectral transform,” JASA, vol. 89, no. 1, pp. 425–434, 1991.

[38] Ealey, Douglas, Holly Kelleher, and David Pearce. "Harmonic tunnelling: tracking non-stationary noises during speech." In *Proc. Eurospeech*, vol. 1, pp. 437-450. 2001.