A New Frequency Coverage Metric and A New Subband Encoding Model, With An Application In Pitch Estimation

Shoufeng Lin

School of Electrical Engineering, Computing and Mathematical Sciences, Curtin University
Kent Street, Bentley, Western Australia, 6102

shoufeng.lin@postgrad.curtin.edu.au; ee.linsf@gmail.com

Abstract

The auditory filterbank has been a well-accepted and important tool for speech feature extraction. It decomposes the speech signal into subbands usually on an equivalent rectangular bandwidth frequency scale before further subband analysis and processing, such as auto-correlation and cross-correlation. However, the choice of the number of subbands and subband center frequencies for a given frequency range has been essentially empirical in the literature. Moreover, correlation of subband signals may not produce distinct peaks of coefficients for feature extraction. This paper proposes a novel frequency coverage metric to calculate the required number of subbands for a given frequency range. It also presents a new subband encoding model for correlation processing, inspired by psychoacoustic studies and statistical analysis. The proposed frequency coverage metric and the subband encoding model are applied to a pitch estimation method as an example of their possible implementations in the speech feature extraction. Compared with state-of-the-art methods, evaluation results demonstrate the benefits of the proposed methods.

Index Terms: auditory filterbank, frequency coverage, subband encoding, pitch estimation, speech feature extraction, CASA.

1. Introduction

Subband decomposition using an auditory filterbank has been a popular method in speech signal analysis. It can be applied to a variety of speech feature extraction applications, such as pitch estimation and tracking [1, 2, 3, 4], speaker localization and tracking [1, 5, 6, 7], and speaker recognition [8, 9].

A critical part of the subband approach in speech processing is the selection of center frequencies for subband filters, according to the chosen frequency range. This is usually addressed by choosing a frequency scale and the corresponding number of subbands. Various frequency scales have been used in the pitch estimation literature, including the logarithmic [10], the Bark [11] and the popular equivalent rectangular bandwidth (ERB)-rate scales [12]. However, in the current literature, the number of subbands for a given frequency scale in the given frequency range largely varies from one implementation to another, essentially as empirical choices with no clear mathematical motivations. In [13], a total of 20 subbands are used for frequency range of 330Hz to 3700Hz, while [2] implements 128 gammatone filters between 80Hz and 5000Hz, and [4] used 48 subbands.

After subband decomposition via the auditory filterbank, auto-correlation or cross-correlation operations are often applied to subband signals for speech feature extraction [2, 3, 4, 9]. In fact, the performance can be improved by encoding the subband signals before the correlation operations, and various heuristic encoding models can be found in the literature [5, 6, 7].

This paper proposes in Section 3 a novel frequency coverage metric for the selection of the number of subbands of a filterbank, derived from a generalized form of the ERB-rate scale. Section 4 presents the new subband encoding model inspired by psychoacoustic and statistical studies. An application of the proposed methods in pitch estimation is provided in Section 5. Section 6 and 7 give the numerical studies and conclusions, respectively.

2. Background - Subband Decomposition

Using an auditory filterbank to decompose the speech signal, the resulting subband signal is denoted as

$$x^{(b)}(t) = g^{(b)}(t) * x(t), \qquad (1)$$

where $x(t)$ is the speech signal, $g^{(b)}(t)$ is the time-aligned filter impulse response of subband b, t is time indix, and integer $1 \le b \le N_b$. Integer N_b is the total number of subbands, and $*$ the convolution operator. Common auditory filters include the gammatone filter [14, 15, 8], gammachirp filter, etc. as well as their variants. In this paper, we use the gammatone filter, which can be expressed as

$$g^{(b)}(t) = \tilde{g}^{(b)}(t) \cdot \cos(2\pi f_C^{(b)} t), \qquad (2)$$

where

$$\tilde{g}^{(b)}(t) = (t + t_d)^{\vartheta - 1} e^{-2\pi f_b^{(b)}(t + t_d)}, \qquad (3)$$

integer ϑ is the order of filter ($\vartheta = 4$ in this paper), t_d is time delay for alignment between filter bands [15], $f_b^{(b)}$ scaling factor for the bandwidth [14, 8], and $f_C^{(b)}$ is the center frequency of filter band b.

3. A New Frequency Coverage Metric

To obtain the subband center frequency $f_C^{(b)}$ for a given audible frequency range $[f_{min}, f_{max}]$, the ERB-rate scale (ERBS) as developed in [12] is often applied, and equi-distant frequencies on the ERBS are then selected based on an empirically chosen number of subbands. Here we propose a metric in order for the derivation of the number of subbands for an arbitrary frequency range.

Denote the general form of ERB as

$$\upsilon(f) = D + E \cdot f, \qquad (4)$$

where $D = 24.7$, and $E = 0.108$ as given in [12], and f the frequency.

From (4), the resulting ERBS becomes:

$$\Upsilon(f) \triangleq \int \frac{1}{\upsilon(f)} df = E' \lg(1 + D' \cdot f), \qquad (5)$$

with the boundary condition $\Upsilon(0) = 0$. Here $D' \triangleq \frac{E}{D}$ and $E' \triangleq \frac{1}{E \cdot \lg e}$. As given in [12], $E' = 21.4$, and $D' = 0.00437$.

The proposed "frequency coverage" metric is defined as

$$\eta_C^{(b)} \triangleq \frac{\frac{1}{2} \cdot (f_B^{(b+1)} + f_B^{(b)})}{f_C^{(b+1)} - f_C^{(b)}}, \tag{6}$$

where $f_B^{(b)}$ denotes the filter bandwidth of subband b. Apparently, a filterbank has consistent and full frequency coverage when $\eta_C^{(b)} \equiv 1$. For $\eta_C^{(b)} < 1$, there are some frequencies falling out of the pass-bands of the filterbank, which may result in estimation error when these frequencies include the desired frequency components. The case of $\eta_C^{(b)} > 1$ still leads to full frequency coverage, but there are redundancies as some frequency components are captured and analyzed multiple times.

The linear relationship between bandwidth and center frequency holds for certain types of filters. Particularly, for the gammatone filter we have [15]:

$$f_B^{(b)} = K_\vartheta \cdot f_b^{(b)} = K_\vartheta \cdot \upsilon(f_C^{(b)}), \tag{7}$$

where K_ϑ is a constant for a given filter order ϑ as given in (8) [15]. In particular, $K_4 = 0.887$.

$$K_\vartheta = 2\sqrt{2^{1/\vartheta} - 1} \cdot \left[\frac{\pi(2\vartheta - 2)! 2^{-(2\vartheta-2)}}{(\vartheta - 1)!^2} \right]^{-1}. \tag{8}$$

The subband center frequencies in the given frequency range are distributed equidistantly on the ERBS, i.e.:

$$f_C^{(b)} = \Upsilon^{-1}\left(\frac{(N_b - b) \cdot \Upsilon(f_{min}) + (b - 1) \cdot \Upsilon(f_{max})}{N_b - 1} \right), \tag{9}$$

where $\Upsilon^{-1}(\cdot)$ is the inverse function of $\Upsilon(\cdot)$.

Therefore the number of subbands N_b can be derived from (6), (7) and (9):

$$N_b = \text{round}\left(1 + \frac{\ln\left(\frac{D+E \cdot f_{max}}{D+E \cdot f_{min}} \right)}{\ln\left(\frac{2\eta_C^{(b)} + E \cdot K_\vartheta}{2\eta_C^{(b)} - E \cdot K_\vartheta} \right)} \right). \tag{10}$$

Thus the frequency coverage metric provides a consistent way for calculating the number of subbands in a given frequency range. Once N_b is obtained, the center frequencies can also be calculated from (9). Since we keep $\eta_C^{(b)}$ the same for all subbands, η_C is used hereafter for simplicity of denotation.

4. A New Subband Encoding Model

After the subband decomposition, we first half-wave rectify the subband signal (1) as in [16, 17, 18].

$$\hat{x}^{(b)}(k/f_s) = \frac{1}{2} \cdot (x^{(b)}(k/f_s) + |x^{(b)}(k/f_s)|), \tag{11}$$

where the $k \in \mathbb{Z}$ is the discrete time index and $f_s > 0$ the sampling frequency.

From (7), each subband has a narrow passband for frequencies not too low. Thus by approximating the subband signal with sinusoid (see e.g. middle panel of Fig. 1), we can rewrite $\hat{x}^{(b)}(k/f_s)$ in (11) as a convolution of the cosine term with the local peaks (the approximation is reasonable as the signal peaks have the dominant effect on results in correlation operations):

$$\hat{x}^{(b)}(k/f_s) \approx \zeta_{\text{cosine}}^{(b)}(k) * \sum_{\hat{k}_n^{(b)} \in \hat{K}^{(b)}} \tilde{S}^{(b)}(k/f_s) \cdot \delta(k - \hat{k}_n^{(b)}), \tag{12}$$

where $\delta(\cdot)$ is the Dirac delta function, $\tilde{S}^{(b)}(k/f_s)$ is the subband envelope, and $\zeta_{\text{cosine}}^{(b)}(k)$ is the non-negative part of the cosine term with peak at $k = 0$, i.e.

$$\zeta_{\text{cosine}}^{(b)}(k) \triangleq \cos(2\pi f_C^{(b)} \cdot k/f_s), k \in \left[-\frac{f_s}{4f_C^{(b)}}, \frac{f_s}{4f_C^{(b)}} \right], \tag{13}$$

$\hat{K}^{(b)} \triangleq \{\hat{k}_n^{(b)} | n = 0, 1, ...\}$, and $\hat{k}_n^{(b)}$ is the index of a local peak (hence $\tilde{S}^{(b)}(\hat{k}_n^{(b)}/f_s) \equiv x^{(b)}(\hat{k}_n^{(b)}/f_s)$)

$$\hat{k}_n^{(b)} = \arg \max_k \hat{x}^{(b)}(k/f_s), \forall k \in (k_{n-}^{(b)}, k_{n+}^{(b)}), \tag{14}$$

$k_{n-}^{(b)}, k_{n+}^{(b)}$ are consecutive zero-crossings of $\hat{x}^{(b)}(k/f_s)$ that satisfy

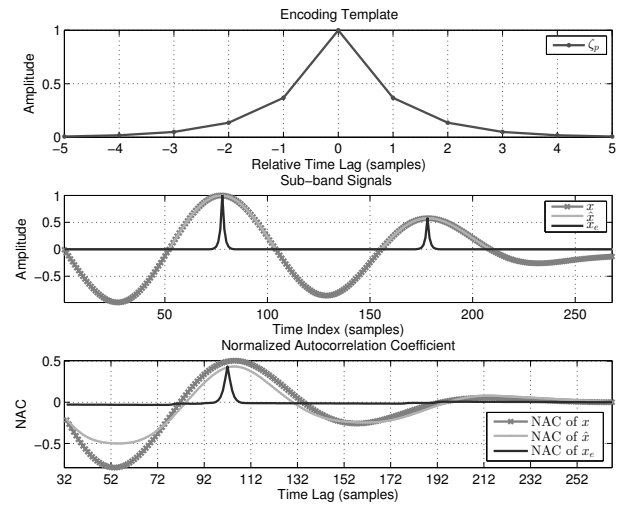$$\hat{x}^{(b)}(k/f_s) > 0, \forall k \in (k_{n-}^{(b)}, k_{n+}^{(b)}). \tag{15}$$

Figure 1: *Subband encoding model (top panel), a subband signal from the filterbank, its half-wave rectified and encoded signal (middle panel), and normalized auto-correlation coefficient of respective signals (bottom panel).*

Usually correlation operations of these subband signals are used for speech feature extraction, but the slow-changing cosine term can make the peak widespread or even cause spurious peaks in correlation coefficients. Taking the pitch estimation for example, apparently the pitch is related to the time intervals between peaks as denoted by the scaled delta functions, i.e. $\tilde{S}^{(b)}(k/f_s) \cdot \delta(k - \hat{k}_n^{(b)})$ as in (12). The problem is that the voiced speech signal is quasi-periodic, and the noise can also offset the time indices of peaks, which affects the correlation coefficients. Therefore, inspired by the approaches of computational auditory scene analysis (CASA) [8, 19], we propose to encode the subband signals as convolution of the scaled delta functions with a symmetrical encoding model (as shown in top panel of Fig. 1), which in effect replaces the cosine term in (12):

$$\zeta_p(k) \triangleq \begin{cases} e^{-|k|}, & k \in [-5, 5] \\ 0, & \text{otherwise,} \end{cases} \tag{16}$$

where the peak decays to 5% of its strength in about 0.2ms at a sampling rate of $f_s = 16000$S/s (i.e. 16000 samples per second). The encoding model aligns with the psychoacoustic observation of the exponential decay of the synaptic cleft contents

from the hair cell in the organ of Corti [19]. It is symmetrical to avoid bias in the correlations. Moreover, the encoding model is also supported by the observation of the Laplacian distribution of the period of peaks versus the deviations [2, 20], except that for simplicity we discard (truncate) smaller values in (16) and the constant coefficient for the exponential term is 1 as it does not affect the resulting normalized correlation coefficients.

The resulting encoded subband signal from (12) and (16) is

$$x_e^{(b)}(k) = \zeta_p(k) * \sum_{\hat{k}_n^{(b)} \in \hat{K}^{(b)}} x^{(b)}(k/f_s) \cdot \delta(k - \hat{k}_n^{(b)}). \quad (17)$$

The top two panels of Fig. 1 depict the encoding model, a segment of a subband signal, its half-wave rectified signal and its encoded signal respectively. Normalized auto-correlation coefficients given in the bottom panel are to be further discussed next in the pitch estimation application.

5. Application in Pitch Estimation

The proposed frequency coverage metric and subband encoding model can be used in various speech feature extraction applications. Here we present a pitch estimation method as an example.

The pitch frequency range is denoted as $[F0_{min}, F0_{max}]$. In this paper, we choose $F0_{min} = 60$Hz, and $F0_{max} = 500$Hz to cover the pitch range of most speakers [21, 22]. Accordingly, the minimum subband frequency is chosen as $f_{min} = F0_{min} = 60$Hz. It has been pointed out that while low frequency auditory nerve fibers of inner hair cells tend to phase lock to pitch stimulus, those of frequencies above 1300Hz do not [13]. Thus we choose $f_{max} = 1270$Hz in this paper [13]. Hence for $\eta_C = 1$ we can get $N_b = 18$ from (10) for the frequency range of $[60, 1270]$Hz.

The encoded subband signals are further processed via auto-correlation in frames of length $n_{corr} = \lceil 2 \cdot f_s/F0_{min} \rceil$ and in step size of $n_{step} \in \mathbb{N}$. The range of sample delays is $d_\tau \in [d_{min}, d_{max}]$, where $d_{min} = \lfloor f_s/F0_{max} \rfloor$, $d_{max} = \lceil f_s/F0_{min} \rceil$. Here $\lfloor \cdot \rfloor$ denotes the largest integer less than or equal to a given number, while $\lceil \cdot \rceil$ denotes the smallest integer greater than or equal to a given number.

Normalized auto-correlation coefficients (NAC) for encoded subband b in the jth frame can be calculated using

$$A^{(b)}(j, d_\tau) = \frac{\sum_{k=(j-1) \cdot n_{step}+1}^{(j-1) \cdot n_{step}+n_{corr}-d_\tau} \tilde{x}_e^{(b)}(k) \cdot \tilde{x}_e^{(b)}(k+d_\tau)}{\sum_{k=(j-1) \cdot n_{step}+1}^{(j-1) \cdot n_{step}+n_{corr}} [\tilde{x}_e^{(b)}(k)]^2},$$

$$(18)$$

where $\tilde{x}_e^{(b)}(k)$ is $x_e^{(b)}(k)$ with dc offset removed for the normalization.

In each time frame, we use the average of the NAC over subbands, i.e. $A_\sum(j, d_\tau) = \frac{1}{N_b} \sum_{b=1}^{N_b} A^{(b)}(j, d_\tau)$ to obtain the correlogram. Then the pitch estimate in time frame j is $\widehat{F_0}\{j\} = \{f_s/\hat{d}_\tau\}$, where $\hat{d}_\tau$ is the sample delay that corresponds to the peaks in $A_\sum(j, \cdot)$ and $A_\sum(j, \hat{d}_\tau) \geq T_{A_\sum}$. If $\hat{d}_\tau$ does not exist, $\widehat{F_0}\{j\} = \emptyset$. The strongest peak over the threshold $T_{A_\sum} = 0.125$ (i.e. -9dB) is used as the pitch estimate.

To show the effectiveness of the proposed encoding model, Fig. 2 provides a single pitch example comparing the proposed estimator using the encoded subband signals (17) and a reference method using half-wave rectified subband signals (11). The top row provides the resulted pitch estimation results using the proposed method and the reference method. We can see that
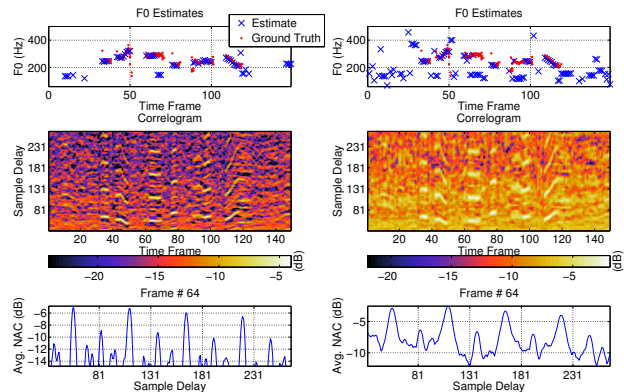

Figure 2: *Pitch estimation results (female speech with babble noise, SNR=10dB). Left column gives the pitch estimation results from proposed method. Right column shows the results using the half-wave rectified subband signals.*

the proposed method produces more valid estimates, while the reference method produces considerably more errors. The middle row depicts the correlogram from the proposed method and the reference method. The proposed method produces more distinct pitch patterns. The bottom row shows the averaged auto-correlation results at frame 64, where the proposed method correctly produces the pitch estimate, while the reference method produces a sub-harmonic error. Therefore it is clear that the proposed method has distinct peaks by virtue of the proposed pitch encoding, while the peaks of the reference method are comparatively widespread. Moreover, using the raw half-wave rectified subband signals produces more sub-harmonics errors. For both cases, spurious estimates when there are no voiced sounds in the ground truth speech signal are due to the babble noise.

6. Numerical Studies

In this section, we compare the pitch estimation results using the proposed methods with other state-of-the-art pitch estimators, namely the RAPT [23], YIN [24], PEFAC [25] and SHRP [10], in noisy environments at various levels of signal-to-noise ratio (SNR). We use the CSTR corpus [26, 27], which includes 50 English sentences from a male and a female speaker respectively and their corresponding pitch ground truth (laryngograph signal). The noise signals used are from the AURORA database [28, 29], which are composed of 8 types of noises from different environments.

6.1. Performance Metric

We use the standard gross pitch error (GPE) [20, 30] for evaluating the performance $\text{GPE} = \frac{N_{err}}{N_v}$, where N_{err} is the number of frames with pitch estimates that deviate from ground truth by more than 5%, and N_v denotes the total number of voiced frames as reported by both the ground truth and the estimation method.

6.2. Test Results

In Fig. 3, we show the GPE results for all the methods with various types of noise and SNR levels. Here $\eta_C = 1$ is used. The parameters for RAPT, YIN, PEFAC and SHRP are the default values as provided in respective programs (hence might not be the most optimal). We choose a frame length of 33.3ms for our
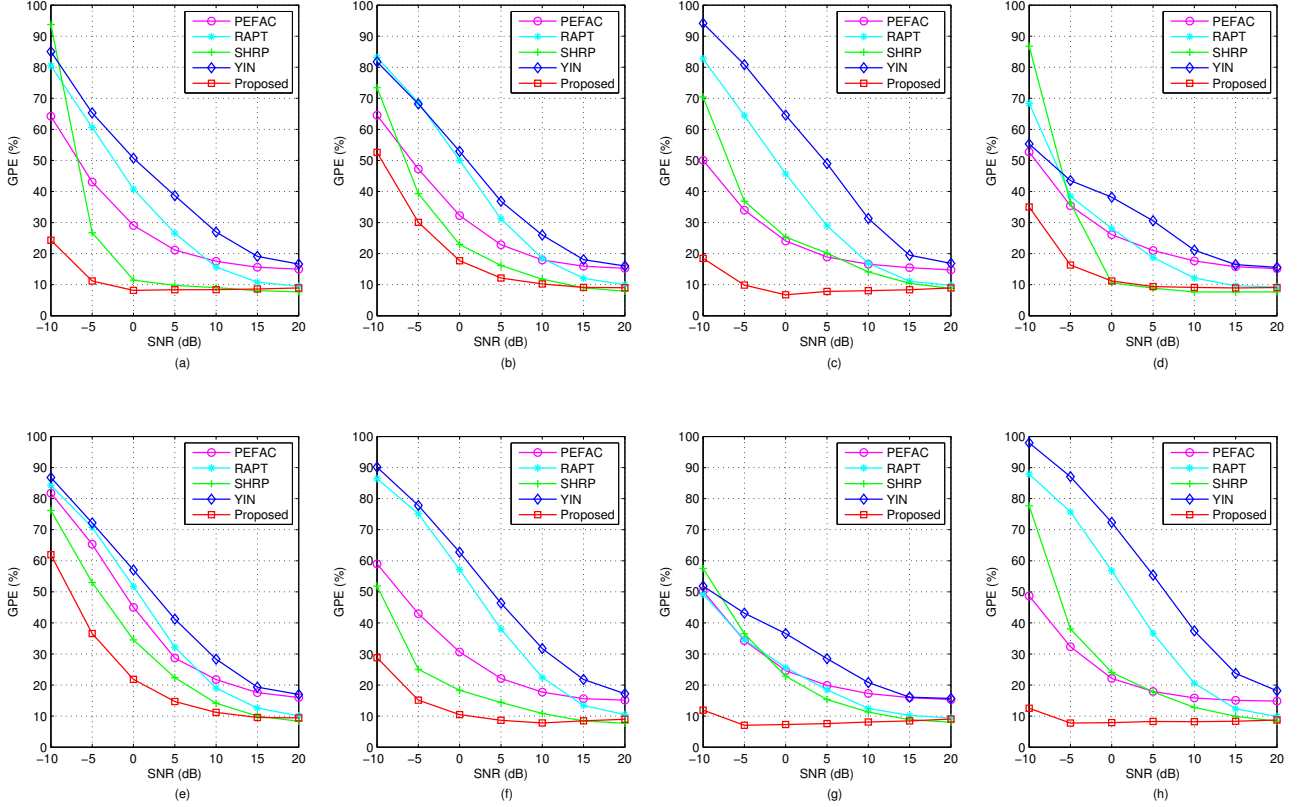
Figure 3: *GPE results for pitch estimation. Speech signals are from the CSTR database, while the additive noises are from the Aurora database. Noise types are respectively (a) airport, (b) babble, (c) car, (d) exhibition, (e) restaurant, (f) street, (g) subway, (h) train.*

proposed method, which is two periods of the minimum F0 frequency ($F0_{min} = 60$Hz). We can see from Fig. 3 that all methods degrade as the noise get stronger. However, the proposed method outperforms the other state-of-the-art methods in most cases. This can be mainly attributed to the proposed subband encoding method as well as the NAC as discussed in Sections 4 and 5 respectively. The performance of the proposed method is worst at the babble noise or the restaurant noise, both of which are basically random mixtures of human speech signals.

Fig. 4 further provides the evaluation results of the proposed frequency coverage metric, where we show the impact of the proposed frequency coverage η_C on the GPE results for the male and female speakers of the CSTR corpus. Here the sound signals are the same as used in Fig. 3, but the GPE is an averaged result over all noise types. We can clearly see that despite the changes of SNR, the accuracy improves (the gross pitch error decreases) as η_C increases until $\eta_C = 1$, and the GPE is comparatively stable for $\eta_C \in [1, 1.5]$. We know that as η_C increases, the number of subbands also increases, thus requiring more computations. Hence as we have expected in Section 3, $\eta_C = 1$ can be chosen for a good balance of the estimation accuracy and the computational load.

7. Conclusions and Future Studies

This paper presents a novel frequency coverage metric for the selection of the number of subbands and subband center frequencies for an auditory filterbank, a new subband encoding model, and an implementation of the proposed methods in the pitch estimation. Comparative study versus the state-of-the-arts
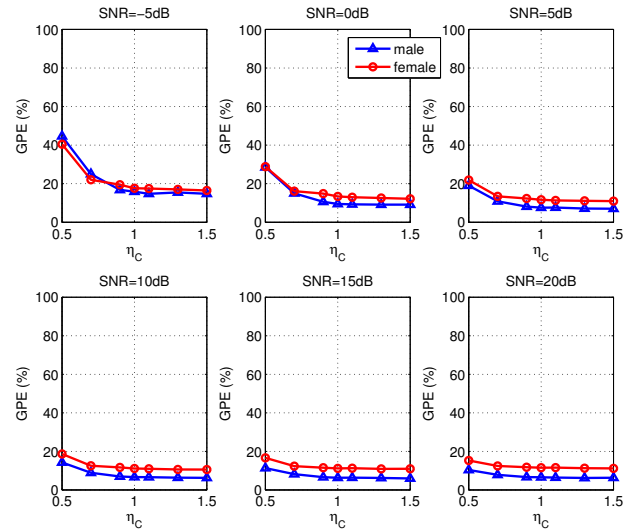

Figure 4: *GPE versus frequency coverage η_C (averaged over all noise types).*

methods shows the benefits of the proposed methods. For future work, other possible implementations of the proposed methods in speech feature extraction include, but are not limited to, the pitch tracking, speaker localization, speech separation, recognition and transcription.

8. References

[1] S. Lin, "Jointly tracking and separating speech sources using multiple features and the generalized labeled multi-bernoulli framework," in *IEEE International Conference on Acoustics, Speech and Signal Processing, 2018. ICASSP 2018. Accepted.*

[2] M. Wu, D. Wang, and G. J. Brown, "A multipitch tracking algorithm for noisy speech," *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 3, pp. 229–241, 2003.

[3] A. Klapuri, "Multipitch analysis of polyphonic music and speech signals using an auditory model," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 2, pp. 255–266, 2008.

[4] B. S. Lee and D. P. Ellis, "Noise robust pitch tracking by subband autocorrelation classification," in *Proc. INTERSPEECH, Portland, OR, USA*, Sep, 2012.

[5] L. S. Smith and S. Collins, "Determining itds using two microphones on a flat panel during onset intervals with a biologically inspired spike-based technique," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 8, pp. 2278–2286, 2007.

[6] A. Plinge, M. H. Hennecke, and G. A. Fink, "Robust neuro-fuzzy speaker localization using a circular microphone array," in *Proc. Int. Workshop on Acoustic Echo and Noise Control, Tel Aviv, Israel*. Citeseer, 2010.

[7] A. Plinge and G. A. Fink, "Online multi-speaker tracking using multiple microphone arrays informed by auditory scene analysis," in *Signal Processing Conference (EUSIPCO), 2013 Proceedings of the 21st European*. IEEE, 2013, pp. 1–5.

[8] D. Wang and G. J. Brown, *Computational auditory scene analysis: Principles, algorithms, and applications*. Wiley-IEEE Press, 2006.

[9] Y. Shao, S. Srinivasan, Z. Jin, and D. Wang, "A computational auditory scene analysis system for speech segregation and robust speech recognition," *Computer Speech & Language*, vol. 24, no. 1, pp. 77–93, 2010.

[10] X. Sun, "Pitch determination and voice quality analysis using subharmonic-to-harmonic ratio," in *2002 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 1. IEEE, 2002, pp. I–333.

[11] J. O. Smith III and J. S. Abel, "Bark and erb bilinear transforms," *IEEE Transactions on Speech and Audio Processing*, vol. 7, no. 6, pp. 697–708, 1999.

[12] B. R. Glasberg and B. C. Moore, "Derivation of auditory filter shapes from notched-noise data," *Hearing research*, vol. 47, no. 1, pp. 103–138, 1990.

[13] J. Rouat, Y. C. Liu, and D. Morissette, "A pitch determination and voiced/unvoiced decision algorithm for noisy speech," *Speech Communication*, vol. 21, no. 3, pp. 191–207, 1997.

[14] R. Patterson, I. Nimmo-Smith, J. Holdsworth, and P. Rice, "An efficient auditory filterbank based on the gammatone function," in *a meeting of the IOC Speech Group on Auditory Modelling at RSRE*, vol. 2, no. 7, 1987.

[15] J. Holdsworth, I. Nimmo-Smith, R. Patterson, and P. Rice, "Implementing a gammatone filter bank," *Annex C of the SVOS Final Report: Part A: The Auditory Filterbank*, vol. 1, pp. 1–5, 1988.

[16] R. Lyon, "A computational model of binaural localization and separation," in *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP'83.*, vol. 8. IEEE, 1983, pp. 1148–1151.

[17] R. Meddis and L. OMard, "A unitary model of pitch perception," *The Journal of the Acoustical Society of America*, vol. 102, no. 3, pp. 1811–1820, 1997.

[18] T. Tolonen and M. Karjalainen, "A computationally efficient multipitch analysis model," *IEEE transactions on speech and audio processing*, vol. 8, no. 6, pp. 708–716, 2000.

[19] R. Meddis, "Simulation of mechanical to neural transduction in the auditory receptor," *The Journal of the Acoustical Society of America*, vol. 79, no. 3, pp. 702–711, 1986.

[20] W. Chu and A. Alwan, "Safe: A statistical approach to f0 estimation under clean and noisy conditions," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 3, pp. 933–944, 2012.

[21] J. R. Deller Jr, J. G. Proakis, and J. H. Hansen, *Discrete time processing of speech signals*. Prentice Hall PTR, 1993.

[22] F. Nolan, "Intonational equivalence: an experimental evaluation of pitch scales," in *Proceedings of the 15th International Congress of Phonetic Sciences, Barcelona*, vol. 39, 2003.

[23] D. Talkin, "A robust algorithm for pitch tracking (rapt)," *Speech coding and synthesis*, vol. 495, p. 518, 1995.

[24] A. De Cheveigné and H. Kawahara, "Yin, a fundamental frequency estimator for speech and music," *The Journal of the Acoustical Society of America*, vol. 111, no. 4, pp. 1917–1930, 2002.

[25] 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.

[26] P. C. Bagshaw, S. M. Hiller, and M. A. Jack, "Enhanced pitch tracking and the processing of f0 contours for computer aided intonation teaching." *Proc. Eurospeech*, pp. 1003–1006, 1993.

[27] Available at http://www.cstr.ed.ac.uk/research/projects/fda/.

[28] 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.

[29] Available at https://www.ee.columbia.edu/~dpwe/sounds/noise/.

[30] D. Wang, C. Yu, and J. H. Hansen, "Robust harmonic features for classification-based pitch estimation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 5, pp. 952–964, 2017.