Vowel Formant Extraction for Speech Decoding using Filter Banks

Alexandra Mikhael
UCSD
San Diego, CA
amikhael@ucsd.edu

Manzhen Jing
UCSD
San Diego, CA
mjing@ucsd.edu

Abstract

Vowel recognition is the key to develop a full closed loop speech prosthesis to help people with difficulty speaking regain their speech. Each vowel can be uniquely identified by its formants (F1,F2 ...) which are peaks in the spectral envelope of the given speech signal. However, a noise robust and reliable algorithm should be in place to distinguish between the different vowels, varying pitch and speaking style of the subject (native speaker, ...).

This paper introduces Wavelet Transforms to denoise the signal before passing it into a Linear Prediction Coefficient (LPC) based formant extractor to decode the vowels. We also investigate Mel filter banks and compare it to LPC for formant extraction accuracy.

1. Introduction

Speech recognition involves taking speech patterns from a person's voice, processing them through a computer, and then identifying the speech's content. It is interdisciplinary and draws on a variety of disciplines, including artificial intelligence, phonetics, linguistics, signal processing, and information theory.

Speech recognition is widely implicated in our technological life, such as iOS Siri, Google, Amazon Alexa, etc. However, due to the varieties of language dialogs and human speech accents, challenges are being faced in the development of speech prosthetics.

Many academics have worked on voice recognition technologies during the last few decades. Over the years of research and trials, vowel recognition has been proved to be one of the most efficient speech recognition methods, since vowel's formant frequencies are the most identifiable in spectrogram observations [1].

1.1. Theory

The following phases can be used to subdivide the speaking process. First, the speaker's brain develops the neural impulses necessary for producing the intended speech or language information. Second, it is translated into language by the human brain, and the speaker then expresses the thought using a variety of volume, pitch, and timbre. After the speech is delivered, other listeners will be able to hear the sounds made by the speakers. Scientists can use various signal processing and statistical techniques to extract the speaker's speech attributes, like formants, and analyze its content.

In human speech, formants are defined as the resonant frequencies of the vocal tract, which can be seen as peaks in the spectrum. Due to the uniqueness of vowel's spectrum distribution, we can determine and identify different vowels based on their frequencies of their formants.

2. Literature Review and Problem Statement

A vowel extraction system is built with two important components, classification and formant extraction. Formant extraction is the key step to visualize the behaviors of different vowels. In current research studies, the two most dominant approaches are Linear Prediction Coefficient (LPC) based formant extraction, and Mel-frequency cepstral coefficients (MFCC) based formant extraction. [2]

Speech is a non-stationary signal where its characteristics are not constant over time but rather vary depending on the speaker's vocal tract. Short-Time Fourier Transform (STFT) has been widely used in literature to study speech in the frequency domain. However, it has been shown that this method is not reliable and robust in extracting speech characteristics because the STFT assumes that the input signal is stationary. [3] Wavelet transform uses a multiresolution approach, whereas STFT has a fixed time-frequency resolution. The capability to conduct local analysis is one benefit of wavelet transform analysis. The signal appearance that other analysis methods miss, such as breakdown points, discontinuities, etc. [4]

Because of its multi-resolution capabilities (both in time and frequency), Wavelet Transform has gained in popularity when analyzing speech signals, especially for denoising and pre-processing. [5] Engineering research in the field of voice denoising focuses on techniques for restoring the original speech from noisy signals that have been distorted by various disturbances. There are many different types of noise that are prevalent in the environment, including white noise, pink noise, babbling noise, and others. Researchers have become more interested in speech processing over the past few decades in the area of noise removal from voice signals. Typically, wavelet techniques are employed for speech denoising.

Linear prediction coding (LPC) is a time-domain based signal-source modeling, which makes it a powerful tool to characterize vowels behavior during the speech signal.[6] Even though LPC approach is very popular in current research studies, the disadvantage of LPC approach is its lack of ability to analyze local events accurately since LPC processes signals in each frame based on the assumption that human speech is a stationary signal. [7][8]

The MFCC approach, on the other hand, can work as a non-linear transformation to generate human auditory filter bank system. [9] Due to its low complicity in calculation, MFCC system perform formant extraction accurately in a clean (no noise) vowel speech environment, however the strict requirement for speech environment is also the biggest weakness of the MFCC system. [10] Therefore, this also leads scientist to dig deeper into the classification part of the system, where more reliable de-noise methods are needed in order to further benefit the vowel extraction accuracy.

3. Proposed Approach

In this work, we propose a novel approach to extract the formants of vowels for speech decoding. We prose to combine the most efficient methods of signal extraction. We will compare the MFCC method and LPC method by combining them with signal classification method and denoising using wavelets to extract vowel formants.

3.1. Dataset Description

For this work, the dataset includes a male subject saying two different vowels. Each trial, the subject is asked to say a specific vowel and his voice is recorded and sampled at 48 KHz. We have two datasets, one for the vowel /e/ and one for the vowel /ae/, and each of them contain 4 trials. The length of vocalization is not the same for different trials, however, the subject is the same and the vowel produced is the same across all trials for a specific vowel. Formant values and pitch are specific to every individual. Since the individual is the same across all trials, we observe that the pitch is constant across trials, and we expect that the formant values extracted for a specific vowel should not vary much between trials.

3.2. Audio pre-processing

Speech being non-stationary and having high redundancy of samples during silent periods (non-speaking), pre-processing the speech signal is crucial to extracting meaningful information and reducing algorithm processing time.

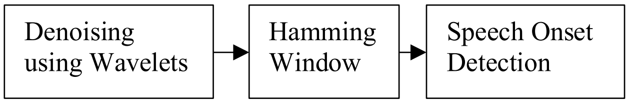


Figure 1: Pre-processing steps diagram

3.2.1 Denoising using Wavelet Transform

The underlying principle of wavelets is scale-based analysis. Wavelet transformations can divide a signal into scales that correspond to various frequency bands, and it is possible to roughly establish the position of the instantaneous structures of the signal at each scale. A signal can be denoised and its time and frequency components examined using such a property. By reducing the wavelet coefficients in the wavelet domain, the wavelet split coefficient approach used in this paper is a speech denoising technique. Each wavelet coefficient in the signal is compared to a predetermined threshold as part of the process. If the coefficient is smaller than the threshold, it is set to zero; otherwise, it is kept or has a tiny amplitude reduction thus the signal is denoised using global thresholding.

The Discrete Wavelet Transform (DWT) used in this work, reduces calculation time while providing sufficient information for analysis and synthesis of speech. It breaks down the signal into a reasonable estimate and detail information after analyzing the signal at various frequency bands and resolutions. The frequency resolution of the human ear is greater at low frequencies and worse at high frequencies. By running the time domain signal through low pass and high pass filters, the signal is decomposed and then analyzed.

By using this method, with a decomposition level of three, we achieve a final signal to noise ratio (SNR) of 2.601 compared to the SNR of the input audio signal that is 12.0118.

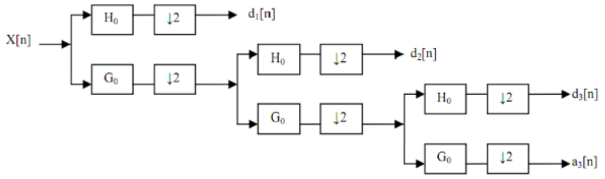


Figure 2: Three Level Wavelet Decomposition Tree

The mother wavelet that we used is the Debauchies wavelet of order 16, where we saw the best results in terms of denoising and robustness across all types of noise introduced. Figure 3 shows an example of the speech signal where we introduced pink noise and figure 4 shows the denoised signal after wavelets.

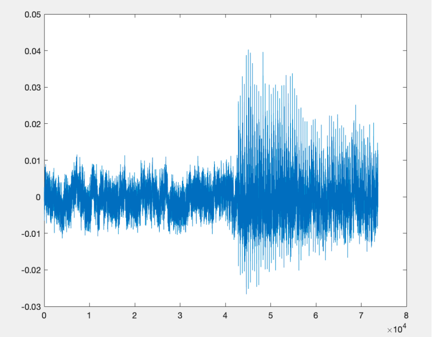


Figure 3: Speech signal before Wavelet denoising

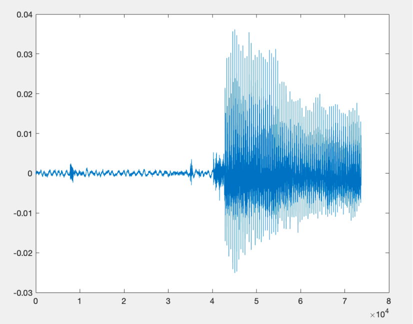


Figure 4: Speech signal after Wavelet denoising

3.2.2 *Hamming Window*

Speech is a non-stationary signal, which means that it varies with time and cannot be considered as linear time invariant. However, it has been shown that a frame of 20 - 30 ms could be considered as time invariant for analysis [11]. For the purpose of this work, we compared two windowing methods : Hamming window and Hanning window. The Hamming window [12] can be described as follows :

$$w(n) = \begin{cases} 0.54 - 0.46 * cos\dfrac{2\pi n}{N-1}, 0 \leq n \leq N-1 & (1) \\ 0, otherwise \end{cases}$$

The Hanning window can be described as :

$$w(n) = \begin{cases} 0.5 - 0.5 * cos\dfrac{2\pi n}{N-1}, 0 \leq n \leq N-1 & (2) \\ 0, otherwise \end{cases}$$

The length of the window is crucial for frequency analysis since if the window is large in the time domain, it will be narrow in the frequency domain and vice versa.

Our window length is 10 ms and the comparison between the two windowing options is shown in figure 5 and figure 6.

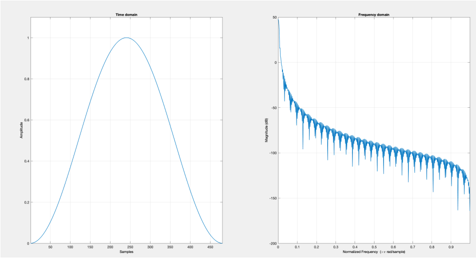


Figure 5: Hamming window

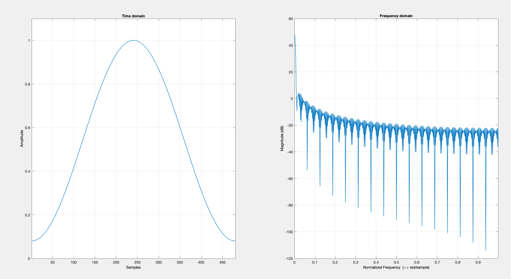


Figure 6: Hanning window

Both windows have the characteristics of low pass and symmetry. The main lobe of hamming window is the widest, has the lowest side lobe level, has relatively stable spectrum for speech signal, and it helps to enhance the characteristics of the central section of signal.

3.2.3 *Speech Onset Detection*

In order to reduce the processing time and errors in formant extraction, we developed a way to detect the onset of speech production. By doing so, we can run the extraction algorithm only on the voiced part of the input signal (speech part), which will resolve the issue of detecting random formants in the silence part.

The detection algorithm is based on the energy and zero-crossing rate of the input signal. In fact, during speech, the energy of the signal is high compared to periods of silence and by setting a threshold, we can detect the onset of speech production and set a classification algorithm that will detect periods of silence, unvoiced or voiced frames. Once a frame is detected to be voiced (i.e speech), the formant extraction algorithm will then run on this specific frame.

Average energy can be defined as:

$$E = \sum_{m=0}^{N-1} [w(m)x(n-m)]^2 , 0 \leq m \leq N-1 \quad (3)[13]$$

where x(n) is the speech signal, N the length of frame, m is the frame shift, w(m) is the window function which expressed as w(m) = $\begin{cases} 1, & 0 \leq m \leq N-1 \\ & 0, otherwise \end{cases}$

Another metric that has been used to identify the onset of speech is zero-crossing rate (ZCR). It shows how many times the x-axis is crossed by a frame of voiced signal. One of the most straightforward techniques for time domain speech analysis is zero crossing analysis. ZCR can be defined as [14] :

$$ZCR = \frac{1}{2}\sum_{m=0}^{N-1} |sgn[x(m) - sgn[x(m-1)]|. w(n-m) \quad (4)$$

Here sgn[ ] is the sign function, which defined as sgn[x] = $\begin{cases} 1, x \geq 0 \\ -1, x \leq 0 \end{cases}$

Unvoiced sounds have a larger ZCR than voiced sounds because their energy is more concentrated in the upper frequencies. Therefore, we can discriminate between voiced and unvoiced frames using the ZCR and energy of the signal.

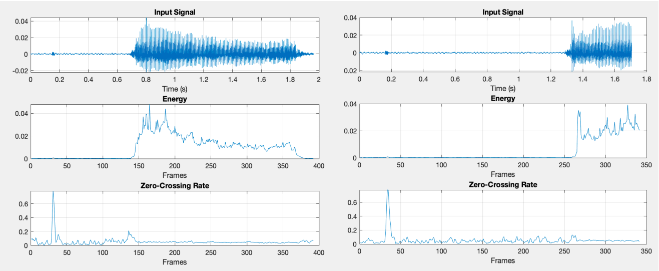


Figure 7 : Energy and Zero-Crossing rate of the two trials

3.3. Algorithm for Formant Extraction

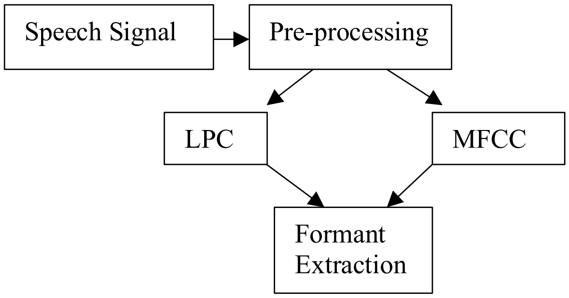
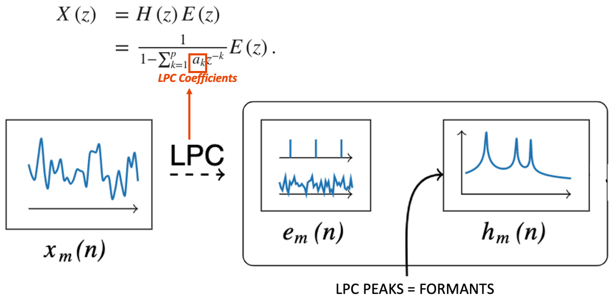


Figure 8: Main Algorithm Flow Diagram

3.3.1 *Linear Prediction Coefficients (LPC)*

One of the most significant and popular speech analysis approaches is linear predictive analysis. This method's significance is rooted in both its relative speed of computation and its capacity to deliver precise estimates of the speech parameters. It is an all pole model in the Z transform domain since the linear predictive analysis method assumes that the speech can be described by a predictor model that only considers previous values of the output.



$$\begin{aligned} X(z) &= H(z)E(z) \\ &= \frac{1}{1-\sum_{k=1}^{p} \boxed{a_k} z^{-k}} E(z)\,. \end{aligned}$$

Figure 9: LPC method diagram

The order of the LPC model is crucial in order to get detailed coefficients of the input signal and be able to accurately extract the frequency components. We compared multiple orders of LPC, ranging from order 3, 15, 30 and 50. The results show that the LPC of order 50 is the one that gives the best approximation of the input speech signal.

After running the algorithm, we get an LPC spectrum for each voiced frame as shown in figure 10.

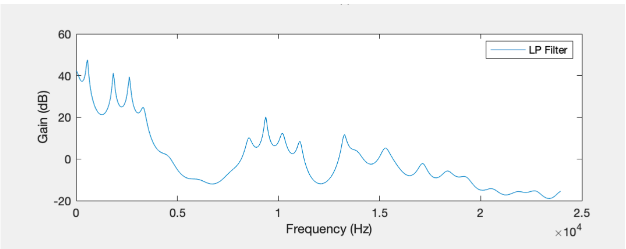


Figure 10: LPC Spectrum for one voiced frame

From this, we can now extract the formants by detecting the peaks in the LPC Spectrum.

3.3.2 *Mel Frequency Cepstral Coefficients (MFCC) Filter Bank*

Based on the mel scale, the method of Mel Frequency Cepstral Coefficients (MFCC) is a powerful methodology. It is important to frame the entire speech signal into a number of sub frames using Hamming windows before calculating the MFCC coefficients. After that, each frame's Fast Fourier transform is calculated. A filter-bank that commonly comprises of overlapping triangle filters that will tailor the frequency resolution to the characteristics of the human ear is used to segment the power spectrum into a number of crucial bands. The raw MFCC vector filters are produced by applying the discrete cosine transformation on the logarithm of the filter-bank outputs. Therefore, the MFCC mimics ear perception behavior and provides a better identification than the LPC. A pure tone's perceived frequency or pitch is related to its actual measured frequency using the Mel scale. Slight pitch changes for low frequency sounds are much easier for humans to distinguish than high frequency sounds. By including this scale, the typical parameters are brought closer to human hearing.

To convert from frequency in Hertz to the mel scale, the following formula is used :

$$M(f) = 1125 * \ln\left(1 + \frac{f}{700}\right) \qquad (5)$$

where f is the frequency in Hertz.

Here, the Mel Filter Bank used consists of 32 bandpass filters in the filter bank with 194 frames in the spectrogram. The Mel Filter Bank function can be visualized in figure 8.

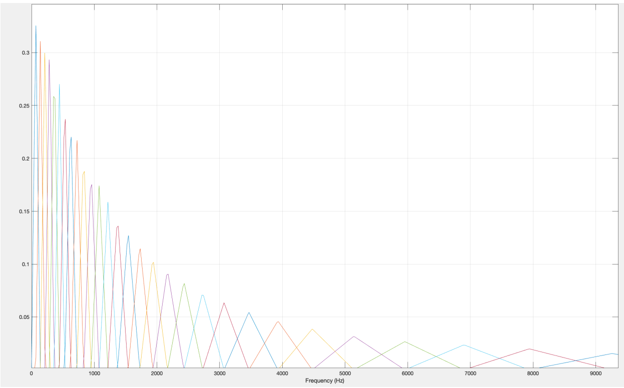


Figure 11: Mel Filter Bank

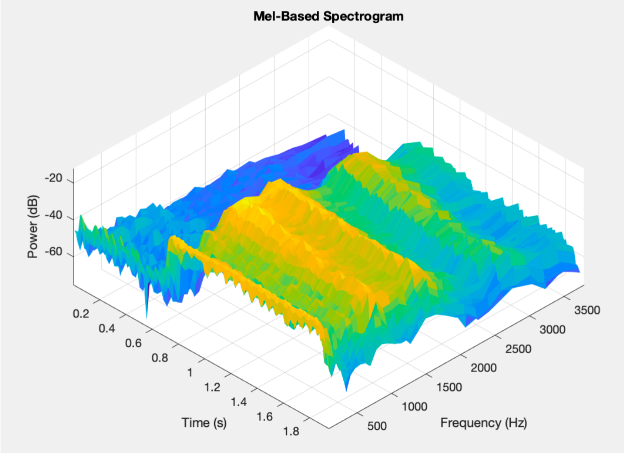


Figure 12: Mel-based Spectrogram for the vowel /ae/

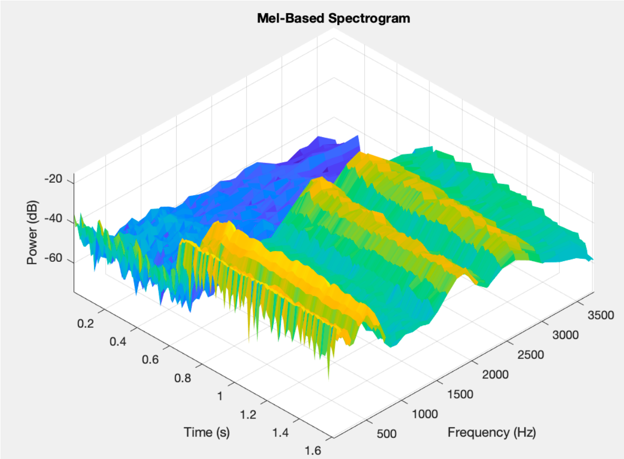


Figure 13: Mel-based Spectrogram for the vowel /e/

We can then detect the peaks in the Mel Spectrograms that will correspond to our formants. We can see from figure 12 and 13 that the spectrograms are different for the two vowels, which is expected.

3.3.3 *Formant Extraction Discussion*

For both methods, LPC and MFCC, the formant values will be the peaks in the envelope of the spectrogram for the MFCC method and the peaks of the LPC spectrum for the LPC method.

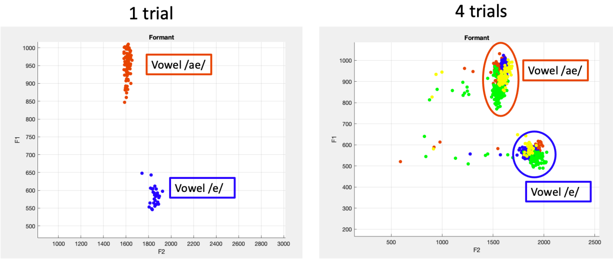


Figure 14: LPC-based Formants extraction

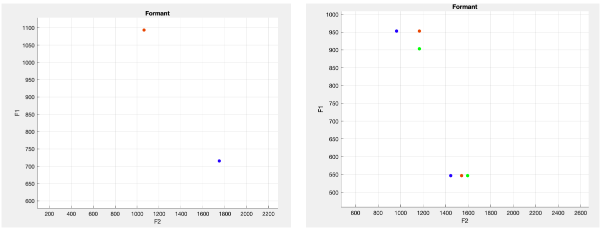


Figure 15: MFCC-based Formants extraction

We can see from the results that both methods give the formant values for F1 and F2 in the same range. However, the LPC method gives a better visualization of the change in F1 and F2 across time in a single trial. In fact, the subject will not hold constant formants over time, and we expect them to vary in a specific range around a given value

(Figure 14). On the other hand, the MFCC method takes the maximum value of the formants throughout the whole period of the trial and so gives us limited temporal resolution where we do not see the formants variations across time (Figure 15). Both methods also show consistency over different trials.

4. Conclusion

In this work, we demonstrated the use of wavelet transforms for speech signal denoising. We also demonstrated the robustness and efficacity of two formant extraction algorithms, MFCC and LPC. These two methods proved to be consistent and accurate in extracting the formants of vowel signals over different trials.

Some differences can be observed between the two. LPC method gives a better temporal visualization of the change in formant value over time in a single trial. However, MFCC method gives us a better visualization of the spectral envelop of the signal over the whole period of the trial. Both methods give formant values in an acceptable range for specific vowels compared to the literature.

In the future, more work should be done in order to create a full-loop speech prosthesis as well as complete sentence decoding.

References

[1] Stam, D. C. (n.d.). “Vowel recognition in continuous speech (thesis)”.

[2] Nehe, N. S., &amp; Holambe, R. S. (2012). DWT and LPC based feature extraction methods for isolated word recognition. EURASIP Journal on Audio, Speech, and Music Processing, 2012(1).

[3] Borisagar, K.R., Thanki, R.M., Sedani, B.S. (2019). Fourier Transform, Short-Time Fourier Transform, and Wavelet Transform. In: Speech Enhancement Techniques for Digital Hearing Aids. Springer, Cham. https://doi.org/10.1007/978-3-319-96821-6_4

[4] Prof. Dr. Ir. M. Steinbuch, Dr. Ir. M.J.G. van de Molengraft, June 7 (2005), Eindhoven University of Technology, Control Systems Technology Group Eindhoven, “Wavelet Theory and Applications”, a literature study, R.J.E. Merry, DCT 2005.53.

[5] Aggarwal, Rajeev, et al. "Noise reduction of speech signal using wavelet transform with modified universal threshold." *International Journal of Computer Applications* 20.5 (2011): 14-19.

[6] Shahriar, S., &amp; Hoq, M. N. (2016). Evaluation of LPC trajectory for Vowel-Consonant-Vowel sequence. 2016 19th International Conference on Computer and Information Technology.

[7] Kent R D and Vorperian H K 2018, “Static measurements of vowel formant frequencies and bandwidths A review”, Journal of communication disorders 74 pp 74-97

[8] Prasad RaviShankar and B. Yegnanarayana, “A study of vowel nasalization using instantaneous spectra” Indian Research Institute, Switzerland. International Institute of Information Technology, India, 2021.

[9] Noelia Alcaraz Meseguer. (2008). Speech Analysis for Automatic Speech Recognition. 87.

[10] Nehe, N. S., &amp; Holambe, R. S. (2012). DWT and LPC based feature extraction methods for isolated word recognition. EURASIP Journal on Audio, Speech, and Music Processing, 2012(1).

[11] D. MANDALIA, P. GARETA, ”Speaker Recognition Using MFCC and Vector Quantization Model,” Institute of Technology, Nirma University, 2011.

[12] Q. Xin och P. Wu, ”Research and Practice on Speaker Recognition Based on GMM,” Computer and Digital Engineering, p. vol 37(6), 2009. Y. Deng, X. Jing, H. Yang,

[13] Y. Yang, ”Research on Endpoint Detection of Speech,” Computer Systems & Applications , p. vol 21(6), 2012.

[14] Y. Cai, ”Research on end point detection of speech signal,” University of Jiangnan, nr TN912.3 Master thesis, 2008.