International Journal of Information, Control and Computer Sciences
ISSN: 2517-9942
Vol:9, No:3, 2015

Speech Enhancement of Vowels Based on Pitch and Formant Frequency

R. Rishma Rodrigo, R. Radhika, M. Vanitha Lakshmi

Abstract**—**Numerous signal processing based speech enhancement systems have been proposed to improve intelligibility in the presence of noise. Traditionally, studies of neural vowel encoding have focused on the representation of formants (peaks in vowel spectra) in the discharge patterns of the population of auditory-nerve (AN) fibers. A method is presented for recording high-frequency speech components into a low-frequency region, to increase audibility for hearing loss listeners. The purpose of the paper is to enhance the formant of the speech based on the Kaiser window. The pitch and formant of the signal is based on the auto correlation, zero crossing and magnitude difference function. The formant enhancement stage aims to restore the representation of formants at the level of the midbrain. A MATLAB software's are used for the implementation of the system with low complexity is developed.

Keywords**—**Formant estimation, formant enhancement, pitch detection, speech analysis.

I. Introduction

SPEECH is a form of communication in every-day life. It existed since human civilizations began and even till now, speech is applied to high technological telecommunication systems. A speech signal is introduced into a medium by a vibrating object as vocal folds in throat. This is the source of the disturbance that moves through the medium. Each spoken word is created using the phonetic combination of a set of vowel semivowel and consonant speech sound units.

Speech sounds are broadly classified into two categories: vowels and consonants. A vowel that is present in the word has higher intensity and longer in duration and also a stronger periodicity whereas consonants have lower intensity and shorter major role compared to consonants. A formant is a concentration of acoustic energy around a particular frequency in the speech wave [1]. In this paper the pitch F0 formant frequencies F1 & F2 are calculated and to improve vowel and consonants discrimination in listeners with hearing loss. Formant frequency correspond to the spectral peaks of the sound spectrum of the voice [3].

A formant is a concentration of acoustic energy around a particular frequency in the speech wave [2]. A recent vowel-coding hypothesis [15] focuses on neural coding of vowels at the level of the auditory midbrain. Many midbrain neurons are not only tuned to the energy within a narrow range around their best audio frequency or best frequency (BF), but are also tuned to the frequency of amplitude modulations [17]. Enhancing of speech degraded by noise, or noise reduction, is the most important field of speech enhancement, and used for many applications such as mobile phones, VoIP, teleconferencing systems, speech recognition, and hearing aids.

Rishma Rodrigo R is with S.A. Engineering College, Chennai, India as a Post Graduate Scholar (e-mail: rishmarodrigo@gmail.com).

Radhika R is with S.A. Engineering College, Chennai, India as an Assistant Professor (e-mail: radhikaus@yahoo.com).

Vanitha Lakshmi M is with Anna University, Chennai, India as a Research Scholar & with S.A. Engineering College, Chennai, India as an Assistant Professor (e-mail: vanithahitesh08@yahoo.co.in).

II. Methods

Traditionally, vowel and consonants encoding studies have focused on representations of formant cues in the output discharge rates of auditory-nerve [4]. The tongue shape and positioning in the oral cavity do not form the major constriction of air flow during the vowel articulation [6]. The relationship of F1 and F2 to one another can be used to describe the English vowels. Consonants as opposed to vowels are characterized by significant constriction or obstruction in the pharyngeal or oral cavity is shown is represented in Fig. 1. Multidimensional analysis of the perceptual vowel space has ascertained that the two dimensions that account for the most variance in the perceptual space correspond to the first two formant frequencies [12]–[14]. The signal-processing system tracks time-varying formants in voiced segments of the input and increases the dominance of a single harmonic near each formant in order to decrease F0-related fluctuations in that frequency channel.

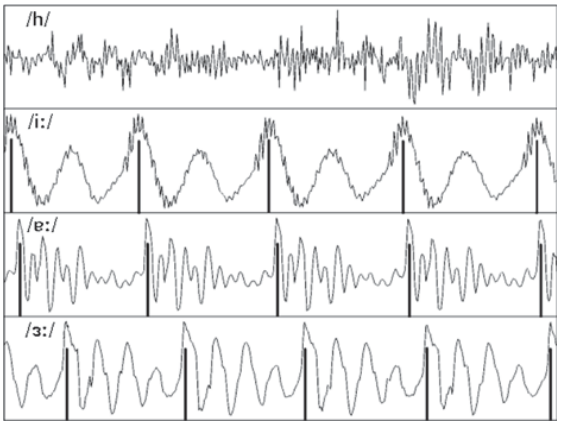

Fig. 1 Spectrogram of some vowel sounds

i). Signal Preprocessing

The signal pre-processing is done in the speech signal to make the signal as a linear one. The signal pre-processing is done. The input speech signal is divided into 32 ms frame with

an overlap of 50% [5]. The sampling frequency is set as 8000 Hz this translated into a frame length of 256 samples. Hanning windowing technique for calculating sequence

$$s(n) = s_m(n) * w(n) \quad (1)$$

$s_m(n)$- input speech frame; w(n)- window sequence.

The window sequence that is used here is a hanning window

$$w(n) = 0.5\left(1 - \cos\frac{2\pi n}{N-1}\right) \quad (2)$$

N-order of the filter; n-number of sequence.

ii). Pitch Estimation

The pitch of the speech signal is estimated using the auto correlation method [7]. Pitch detection algorithm is an algorithm designed to estimate the pitch or fundamental frequency of a quasi-periodic or virtually periodic signals usually a digital recording of speech or a music note or tone. This can be in either frequency domain or in the time domain or in both the domain [8]. It uses the autocorrelation method in which the fundamental frequency is calculated from the voiced speech signal. Autocorrelation function φ(r) from the input frame s(n) is given by

$$\varphi(r) = \frac{1}{N}\sum_{n=0}^{N-1} s(n)s(n+r) \quad (3)$$

r- lag number; n- time of the discrete signal. By this equation the autocorrelation is done Auto Correlation Based on Pitch extraction

$$Crr(\delta) = \frac{\{\sum s(n)+s(n+\delta)\}}{N-\delta} \quad (4)$$

Crr- autocorrelation sequence; s(n)- current frame; Del- lag or delay; N- frame length; Del(p)- maximum.

iii). Formant Frequency Estimation

The frequency and the amplitude of the first three formantsF1, F2 and F3 during all vowel and consonants like segments of the continuous speech. It uses linear prediction spectra and segment parameter to indicate energy and voicing. processing begins at the middle of each high volume voiced segments where the formants are most recently found at laboratories. Formant features can be interpreted as adaptive non uniform samples of the signal spectrum that are located in the response frequency of the vocal tract and normally happen to have higher signal to noise ratio than the other [10].

The number and the position of these frequencies along the frequency axis might differ depending on the phoneme and the positions of the window along with the phoneme. Long with the formant we might use the bandwidth and magnitude of the spectrum in that frequency to encode the properties of the speech and use them in different applications such as speech recognition, speech enhancement, noise reduction, hearing aid adaptive filter [4]. The first three formants of the signal considered as such as F1, F2. From these frequencies it estimated shown is represented in Fig. 2.

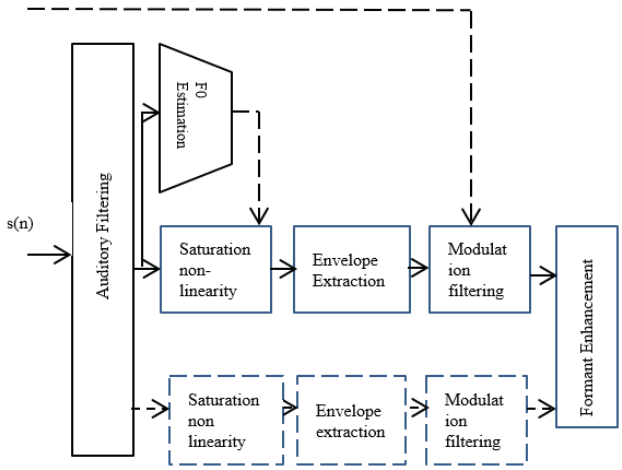

Fig. 2 Schematic of formant estimation.

The inputs to this stage are the speech frame s(n) from signal reprocessing and F0 from the pitch estimation stage. Solid arrows represent flow of the speech signal, while dashed arrows represent flow of parameters such as pitch and formant estimates [7]. Some, but not all, of the corresponding pathways for other channels have been shown using dotted arrows. For the sake of clarity, a few obvious signal paths and operations such as energy criterion and smoothing have been omitted from the schematic.

a) Auditory Filtering

The speech frame decomposed s(n) into multiple band pass channel x(f,n) 77 to 3700Hz increasing the center frequency.

b) Saturation Nonlinearity

It is based on sample by sample analysis which is given by the Sigmoid curve Boltzmann Function [1], [5]

$$xl(f,n) = \frac{A_1 - A_2}{1 + e^{x(f,n)s(n)}} + A_2 \quad (5)$$

$xl(f,n)$- output of non-linearity for band pass-filter channel

c) Modulation Filtering

The filtering of the signal is based on the source spectrum of the signal. Each channel envelope was passed through a narrow band pass filter [9]. Source spectrum threshold old function $S_c(f)$ is nonlinear function of frequency and source filter model

$$S_c(f) = \frac{10^{\frac{-m\log(f/F0)-k}{20}}}{x_{rms}(F0)} \quad (6)$$

f- center frequency of auditory filter channel; c- Index of current frame; F0-Voice pitch of the current frame; $X_{rms}(F0)$- RMS value of the filter output (denoted as F0 channel); m- Source spectrum slope; k- Suitable value of empirically determined (-9dB/octave and 6dB).

III. Formant Enhancement

Mainly two formants are considered the analysis of the signal from which the formant frequencies are calculated and based on this formant frequency values they are calculated. Formant enhancement is based on the formula

$$v1 = \left\lceil \frac{f1}{f0} \right\rceil . f0 \quad (7)$$

$$v2 = \left\lceil \frac{f2}{f0} \right\rceil f0 \quad (8)$$

f1,f2- formant frequencies; f0- pitch of the signal.

The function which is used for analysis of v1 and v0 is called as a ceiling function which is used to round of the function so that greater value of the division is considered. The v1 and v0 are narrowed formant F1 and F2 frequencies [13].

As represented in Fig. 4, first, the frequencies $v1$ and $v2$ of two harmonics were calculated by finding the integer multiples of F0 closest to F1 and F2 [11]. If any formant estimate was found to be equidistant from two adjacent harmonics, the lower harmonic was chosen. Next, two linear-phase narrowband FIR band pass filters, centered at $v1$ and $v2$, respectively, having pass band gains of g1 and g2, amplified the respective harmonics in the current speech frame, s (n). In the current implementation, an FIR filter of order 300 was generated using the Kaiser Window [15] method of FIR filter design, using a bandwidth of 50 Hz and a stop band attenuation of 25 dB. A gain g0 was then applied to the summation in order to account for elevated thresholds in listeners with hearing loss. Appropriate values of these gains would be determined empirically for each subject. The gains g1 and g2 would be fixed across time, and selected based on responses to a range of vowel sounds.

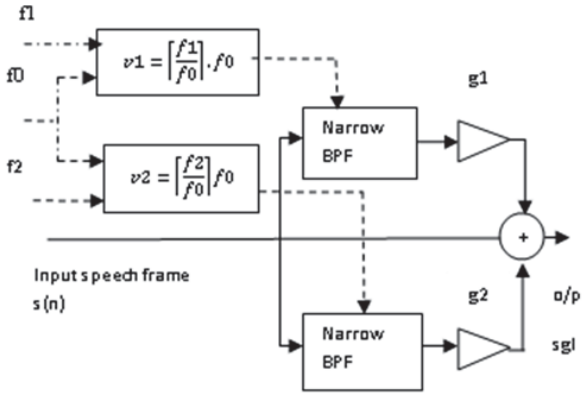

Fig. 3 Schematic diagram of formant Enhancement

IV. Results

A non-real-time implementation of the system with tunable parameters was developed in MATLAB to test the ability of the vowel-coding hypothesis to guide a novel formant-tracking method and to enhance the discrimination of vowels in listeners with hearing loss [16]. In order to compare estimates of the formant-tracking subsystem to the database formant values, the vowel portion from each sample was extracted using the vowel start and end times provided by the database [17]. This segment was then down sampled to 8000 Hz and was passed through the pitch tracking and formant-tracking subsystems. The magnitude of the difference between each formant estimate and its corresponding known formant frequency from the database was normalized using the known F0 value.

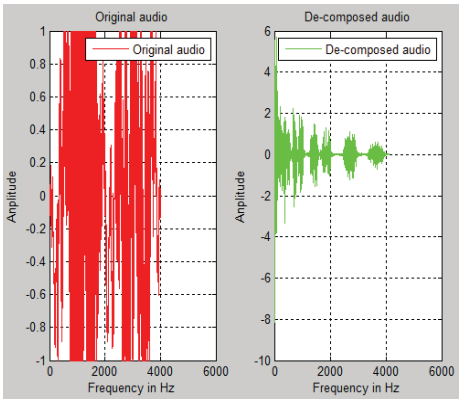

Fig. 4 Decomposition of speech signal

After the decomposition of the speech signal the signal is divided into 256 samples with 32 ms and the pitch of the signal is calculated as such as is shown below in the diagram

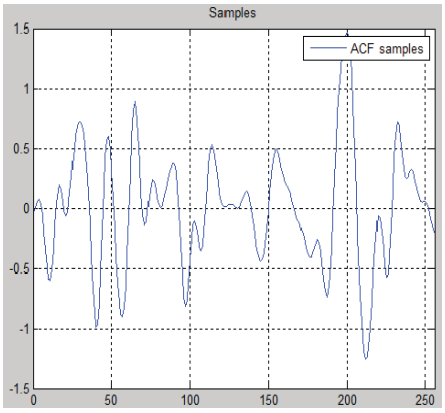

Fig. 5 Autocorrelation function of the signal

The formants f0, f1, f2 of the signal is based on envelope of the speech signal it is done after the separation of the vowels and the noise of the speech signal.

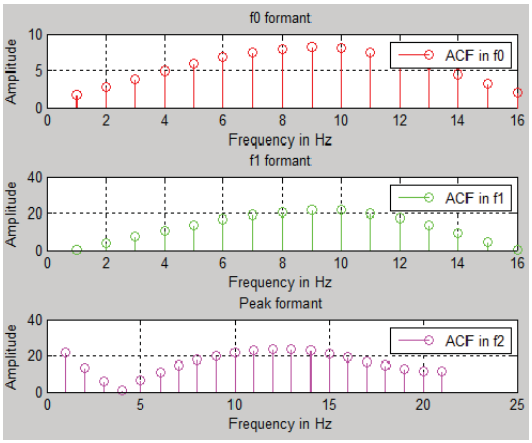

Fig. 6 Peak value of the formants

V. Conclusion

A new strategy has been developed to enhance the speech signal based on the vowels. From these words the consonants and the vowel sounds are identified separately and then from their spectrogram the frequencies are calculated which is further used in the enhancement of the speech signal based on the formant frequencies. The observations also guided the design of a novel formant-tracking strategy, which showed reasonable ability to generalize over multiple speakers' vowels. Objective evaluation of the formant-tracking strategy was carried out and described.

References

[1] Akshay Rao, Laurel H. Carney, "Speech enhancement for hearing loss based on vowel cofing in auditory midbrain" in IEEE transactions on biomedical engineering, vol. 61, no. 7, July 2014.
[2] R. A. Cole, Y. Yan, B. Mak, M. Fanty, and T. Bailey, "The contribution of consonants versus vowels to word recognition in fluent speech," in Proc. IEEE ICASSP, Atlanta, GA, USA, May. 1996, pp. 853–856.
[3] D. Kewley-Port, T. Z. Burkle, and J. H. Lee, "Contribution of consonant versus vowel information to sentence intelligibility for young normal hearing and elderly hearing-impaired listeners," J. Acoust. Soc. Amer., vol. 122, no. 4, pp. 2365–75, Oct. 2007.
[4] M. J. Owren and G. C. Cardillo, "The relative roles of vowels and consonants in discriminating talker identity versus word meaning," J. Acoust. Soc. Amer., vol. 119, no. 3, pp. 1727–39, Mar. 2006.
[5] G. Parikh and P. C. Loizou, "The influence of noise on vowel and consonant cues," J. Acoust. Soc. Amer., vol. 118, no. 6, pp. 3874–3888, Dec. 2005.
[6] G. Fant, Acoustic Theory of Speech Production. Hague, The Netherlands: Mouton, 1960. Rao and Carney: Speech Enhancement for Listeners with Hearing Loss 2091
[7] D. B. Fry, A. S. Abramson, P. D. Eimas, and A. M. Liberman, "The identification and discrimination of synthetic vowels," Language Speech, vol. 5, no. 4, pp. 171–189, Oct.–Dec. 1962.
[8] H. L. Helmholtz, On the Sensations of Tone as a Physiological Basis for the Theory of Music. Cambridge, U.K.: Cambridge Univ. Press, 2009.
[9] M. Ito, J. Tsuchida, and M. Yano, "On the effectiveness of whole spectral shape for vowel perception," J. Acoust. Soc. Amer., vol. 110, no. 2, pp. 1141–1149, Aug. 2001.
[10] S. A. Zahorian and A. J. Jagharghi, "Spectral-shape features versus formants as acoustic correlates for vowels," J. Acoust. Soc. Amer., vol. 94, no. 4, pp. 1966–1982, 1993.
[11] J. D. Miller, "Auditory-perceptual interpretation of the vowel," J. Acoust. Soc. Amer., vol. 85, no. 5, pp. 2114–2134, May. 1989.
[12] R. K. Potter and J. C. Steinberg, "Toward the specification of speech," J. Acoust. Soc. Amer., vol. 22, no. 6, pp. 807–820, 1950.
[13] B. Mohr and W.-Y. Wang, "Perceptual distance and the specification of phonological features," Phonetica, vol. 18, no. 1, pp. 31–45, 1968.
[14] L. C. W. Pols, L. J. T. V. D. Kamp, andR. Plomp, "Perceptual and physical space of vowel sounds," J. Acoust. Soc. Amer., vol. 46, no. 2B, pp. 458– 467, Aug. 1969.
[15] G. Langner and C. E. Schreiner, "Periodicity coding in the inferior colliculus of the cat. I. Neuronal mechanisms," *J. Neurophysiol.*, vol. 60, no. 6, pp. 1799–1822, Dec. 1988.
[16] B. S. Krishna and M. N. Semple, "Auditory temporal processing: responses to sinusoidally amplitude-modulated tones in the inferior colliculus," *J. Neurophysiol.*, vol. 84, no. 1, pp. 255–273, Jul. 2000.
[17] P. C. Nelson and L. H. Carney, "Neural rate and timing cues for detection and discrimination of amplitude-modulated tones in the awake rabbit inferior colliculus," *J. Neurophysiol.*, vol. 97, no. 1, pp. 522–539, Jan.2007.