INTERSPEECH 2020
October 25–29, 2020, Shanghai, China

Enhancing formant information in spectrographic display of speech

B. Yegnanarayana[1], *Anand Joseph*[1,2] *and Vishala Pannala*[1]

[1]Speech Processing Laboratory, IIIT, Hyderabad, India
[2]Sayint Team, TechMahindra
yegna@iiit.ac.in,{anandjm, p.vishala}@research.iiit.ac.in

Abstract

Formants are resonances of the time varying vocal tract system, and their characteristics are reflected in the response of the system for a sequence of impulse-like excitation sequence originated at the glottis. This paper presents a method to enhance the formants information in the display of spectrogram of the speech signal, especially for high pitched voices. It is well known that in the narrowband spectrogram, the presence of pitch harmonics masks the formant information, whereas in the wideband spectrogram, the formant regions are smeared. Using single frequency filtering (SFF) analysis, we show that the wideband equivalent SFF spectrogram can be modified to enhance the formant information in the display by improving the frequency resolution. For this, we obtain two SFF spectrograms by using single frequency filtering of the speech signal at two closely spaced roots on the real axis in the z-plane. The ratio or difference of the two SFF spectrograms is processed to enhance the formant information in the spectrographic display. This will help in tracking rapidly changing formants and in resolving closely spaced formants. The effect is more pronounced in the case of high-pitched voices, like female and children speech.
Keywords: formants, spectrograms, single frequency filtering

1. Introduction

Formants are resonances of the vocal tract system. They are reflected in the response of the system for impulse-like excitation as in most voiced sounds, and sometimes for noise-like excitation as in some fricative sounds. The vocal tract system needs to be excited by a sequence of impulses to sustain a vowel-like sound or to track the dynamic vocal tract shape as in consonant-vowel (CV) or vowel-consonant (VC) sounds. Even though the vocal tract shape changes within a glottal cycle due to source tract interaction, only the resonance frequencies of the supra-glottal vocal tract are generally considered as formants. The changes in the resonance frequencies across successive glottal cycles are called formant contours. Since the formant contours represent the dynamics of the vocal tract system, they are useful for speech analysis to describe the characteristics of different sounds [1, 2, 3]. They are also useful for speech modification [4], speech synthesis, low bit rate speech coding and speech recognition [5].

Methods for formant extraction are developed assuming that the location of peaks in the envelope of the short time spectrum correspond to formants [6]. An estimate of the short time spectrum is obtained by computing the squared magnitude of the discrete Fourier transform (DFT) of a segment of speech signal. The envelope of the short time spectrum is derived using several nonlinear smoothing techniques, the cepstrum-based smoothing being most popular. Smoothed spectrum is also derived using model-based methods, such as linear prediction analysis [7]. By exploiting the high frequency resolution properties of group delay functions, several methods were proposed for extraction of formants from speech signals [8, 9, 10].

Methods for extraction of formant contours involve tracking the formants obtained by peak picking followed by continuity constraints [11]. Other methods use dynamic programming [12], hidden Markov models [13] and Kalman filtering [14, 15]. More recently, deep learning approaches are explored for formant extraction and tracking [16].

Formants and formant contours are mostly visible in the spectrogram display of continuous speech. This is partly due to eye averaging of the formant features. Prior knowledge of the speech and its dynamic spectral characteristics helps in ignoring spurious information and in extrapolating or filling the missing information. The artifacts of signal processing in the computation of the short time spectrum for spectrographic display also affect the visibility of the formant features in the display. Let us discuss briefly some of these artifacts in the context of narrowband and wideband spectrograms. In the narrowband spectrogram the harmonics in the short time spectrum make it difficult to extract the formant peaks from the smoothed envelope of the spectrum. Also, the formant changes in successive pitch periods will not be captured well in the narrowband spectrogram. In the wideband spectrogram the frequency resolution is so poor that it is not only difficult to extract the formant frequencies, but it is also difficult to isolate closely spaced formants. Most of the rapidly changing formant contours will not be visible in the wideband spectrogram. In both types of spectrograms the effect of the size, shape and location of the window with respect to signal can significantly affect the displayed information of formants and formant contours. Smoothing of the short time spectrum for deriving the spectral envelope will further mask the true formant information.

The signal processing artifacts will also affect the formant extraction from model-based spectrum analysis like linear prediction [7]. Since linear prediction analysis approximates the peaks in the short time spectrum, it is likely to give wrong peaks, especially for high pitched sounds. Even in the linear prediction analysis, the rapid changes in the formants cannot be captured well, if the analysis window size spans over several pitch periods. If the window size is less than a pitch period, then the computed autocorrelation coefficients may be biased. Higher order linear prediction analysis produces spurious peaks in the model spectrum. The most important limitation of model-based analysis methods is the need to decide the model order in advance, leading to either wrong peaks or spurious peaks. All these issues were well described in the literature [17].

The signal processing artifacts and the somewhat arbitrary constraints of continuity pose challenges in formant extraction and formant tracking. It is likely that some important information may have been lost in the spectral representation of the signal, either in the short time spectrum analysis or in the model-based analysis. Also, the definition of formants itself may not be clear for a continuously varying vocal tract shape

http://dx.doi.org/10.21437/Interspeech.2020-2653

during speech production. Even within a glottal cycle the shape of the vocal tract changes due to varying degree of coupling of the supraglottal system with the subglottal system in the open and closed phase regions during vocal fold vibrations. The effect of this coupling may result either in shifting of the first resonance to lower frequencies or in increasing the bandwidth of the first formant, or both, as shown in the zero-time windowing method of analysis of speech [18, 19]. In the closed phase of the glottal cycle the resonances may correspond to the supraglottal vocal tract. Since the signal energy is higher soon after the impulse-like excitation occurring at the glottal closure instant (GCI), the signal in the closed phase region of the glottal cycle may dominate in the analysis over a window size of about one pitch period. We assume that the resonances corresponding to the closed phase regions are formants. The dynamic vocal tract system may have fast changes of formant frequencies even from one glottal cycle to the next. Also, there could be abrupt changes in the formant frequencies due to sudden changes in the shape of the vocal tract system. There could be weak presence of some formants due to spectral dynamic range. There may also be presence of resonance frequencies and their tracks in the high frequency regions for unvoiced sounds.

In view of the above discussion, there is need to process the speech signal to obtain at least a display which shows the presence of the formants and formant transitions for all cases of a dynamic vocal tract system. The display of the formant information gives an indication of the dynamic characteristics of the vocal tract system. The objective of the current study is to process the speech signal to display the formant features in continuous speech, rather than extracting them explicitly. This paper proposes a method to display the formant information in spectrogram-like plots. The method overcomes some of the limitations of the block processing methods like the short-time Fourier transform and linear prediction analyses. The method also displays the formant features at every sampling instant due to processing the signal using single frequency filtering (SFF). The effect of harmonics is reduced by analyzing the signal with a tapering window that gives more importance to the current and recent samples. The wideband spectrogram features obtained at each sampling instant are further processed by a novel spectral ratio or difference computation, so that the formant features are displayed with high resolution to resolve closely spaced formants, and also to track abrupt changes and breaks in the formant contours. The results of the proposed method are illustrated for some speech utterances.

The paper is organized as follows. Section 2 describes briefly the short-time Fourier transform (STFT) based and single frequency filtering (SFF) based spectrogram displays, contrasting the methods in displaying the features of the dynamic vocal tract system characteristics during speech production. Section 3 describes a method for highlighting the formant features in the wideband equivalent SFF spectrogram. The issues of temporal and spectral resolution are discussed in some detail. Section 4 discusses the results of the proposed method for some speech utterances. Section 5 gives a summary of the paper with a discussion on the need for extracting the formant information from the SFF spectrograms.

2. Time-frequency representation using spectrograms

In this section, two methods of time-frequency representations of dynamic characteristics of the vocal tract during speech production are reviewed briefly. They are STFT-based and SFF-based spectrum analyses.

2.1. STFT-based spectrum analysis

In this method a short segment of speech signal is considered. The segment is multiplied with a suitable window function such as Hamming or Hann, and the DFT is computed for the windowed segment. The log magnitude of the DFT is the short time spectral representation of the signal around the center of the segment. The segment is also called frame. The short time spectrum can be obtained at every instant by using a frame shift of one sample. The size of the analysis frame can be varied depending on the required resolution of the features in the time and frequency domains. Typically, a window size of 30-50 ms gives a narrowband spectrum analysis, as it gives higher frequency resolution and lower temporal resolution. Since the window size contains several pitch periods, the resulting short-time spectrum will have spectral envelope superimposed by harmonics, making it difficult to locate the formant peaks, especially for high pitched utterances. The log spectrum is displayed in terms of varying grey levels along the vertical frequency axis at each instant of time, resulting in spectrogram display as shown in Fig. 1(b) for the signal in Fig. 1(a). The utterance of the signal is *"She had your dark suit in greasy wash water all year"*, sampled at 8 kHz. The displayed figure is a narrowband spectrogram, as a frame size of 30 ms is used for analysis.

If a frame size of 3 ms (less than a pitch period) is used for analysis, then we get a wideband spectrogram as shown in Fig. 1(c). The dark bands in the spectrograms correspond to the resonances of the vocal tract system, i.e., the formants. Because the frequency resolution is poor, it may be difficult to determine the formant frequencies and to track them. It may be difficult to observe the formant contours clearly in both narrowband and wideband spectrograms. Note that for a given utterance, proper choice of the window size may give significantly better display of formant information than shown in Fig. 1(b) and 1(c).

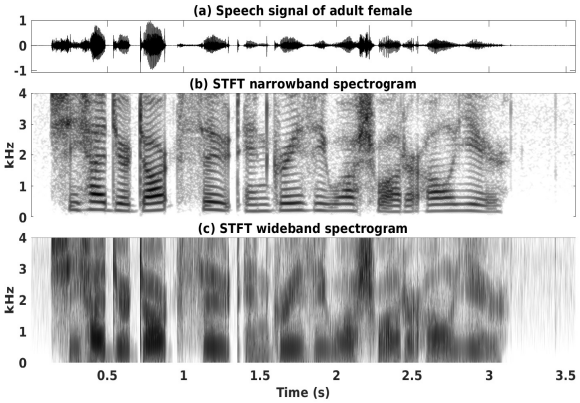

Figure 1: *Illustration of STFT spectrograms for adult female voice. (a) Speech signal. (b) Narrowband spectrogram. (c) Wideband spectrogram.*

2.2. SFF-based spectrum analysis

Single frequency filtering (SFF) is an alternative method for time-frequency analysis of signals [20]. It involves frequency shifting the signal and passing the shifted signal through a near ideal resonator at half the sampling frequency, i.e., $f_s/2$. The filtered signal gives the component of the signal at the desired frequency. The following are the steps involved in the SFF analysis [20].

The signal $s[n]$ is differenced to remove any bias or low frequency fluctuations in the recording. The differenced signal is given by

$$x[n] = s[n] - s[n-1], \qquad n = 1, 2, \ldots, N, \qquad (1)$$

where N is the number of samples in the differenced signal.

The differenced signal $x[n]$ is frequency shifted by multiplying with $e^{j\hat{\omega}_k n}$, where $\hat{\omega}_k = \pi - (2\pi f_k/f_s)$, f_k is the desired frequency in Hz and f_s is the sampling frequency in Hz. The frequency shifted signal $x_k[n]$ is given by

$$x_k[n] = x[n]e^{j\hat{\omega}_k n}. \tag{2}$$

The frequency shifted signal $x_k[n]$ is filtered by a single pole filter $H(z) = 1/(1 + rz^{-1})$, where the pole is located on the negative real axis at $z = -r$ in the z-plane.

The filtered output $y_k[n]$ is given by

$$y_k[n] = -ry_k[n-1] + x_k[n], \qquad n = 1, 2, \ldots, N. \tag{3}$$

The magnitude $e_k[n]$ and phase of $\theta_k[n]$ of $y_k[n]$ are given by

$$e_k[n] = \sqrt{y_{kr}^2[n] + y_{ki}^2[n]}\,, \tag{4}$$

$$\theta_k[n] = tan^{-1}\big(\frac{y_{ki}[n]}{y_{kr}[n]}\big). \tag{5}$$

The $e_k[n]$ for different values of k gives the SFF magnitude spectrum. The number of frequency points K correspond to the number of desired frequency components in the range $0 - f_s/2$. The plot of $e_k[n]$ for all k and n gives the SFF spectrogram. The choice of r controls the frequency and time resolution of the spectral content in the SFF spectrogram. Figs. 2(b) and 2(c) show the SFF spectrograms for $r = 0.995$ and $r = 0.85$, respectively. The SFF spectrogram in Fig. 2(b) appears similar to the narrowband STFT spectrogram in Fig. 1(b). Likewise, the SFF spectrogram in Fig. 2(c) appears similar to the wideband spectrogram in Fig. 1(c). By varying the value of r, different resolutions in time and frequency can be obtained. The objective of this study is to obtain a spectrogram with good spectral and temporal resolutions, so that the formants and formant contours can be observed to relate them to the dynamic characteristics of the vocal tract system. This is done by modifying the wideband equivalent of the SFF spectrogram to improve the frequency resolution of formants.

3. Modification of SFF spectrogram

In this section we propose a method to highlight the formant information in the display of the wideband equivalent SFF spectrogram. The method uses the property that the spectral peaks are sharper in the SFF spectra for higher r, compared to the SFF spectra for relatively lower r. To avoid the effects of harmonics, the SFF spectra are derived using a low value of r, say around $r = 0.85$. This will result in instantaneous SFF spectra with poor spectral resolution, as in the case of wideband spectrogram, although it gives high temporal resolution. Taking the ratio of the SFF spectra for r and $r - \Delta r$, we will notice that the ratio values near the peaks of the SFF spectra will be greater than 1, and the ratio values near the valleys will be less than 1. The SFF ratio spectrum improves the spectral resolution without affecting the temporal resolution, if the value of r is so chosen that the harmonics do not appear in the SFF spectra. We have chosen $r = 0.85$ in this study, although any choice in the range 0.85 to 0.90 seems adequate. Note that the value of r decides the tapering effect of the response of the filter, i.e., the effect of the past samples. If it is too small, then the smoothing along the frequency increases, thus reducing the frequency resolution. The value of r should be low enough to reduce the effects due to previous pitch periods. The SFF ratio spectrum is derived as follows. Let $e_{k1}[n]$ and $e_{k2}[n]$ be the SFF envelopes at r and $r - \Delta r$, respectively. Then the SFF ratio spectrum $e_{kr}[n]$ is given by

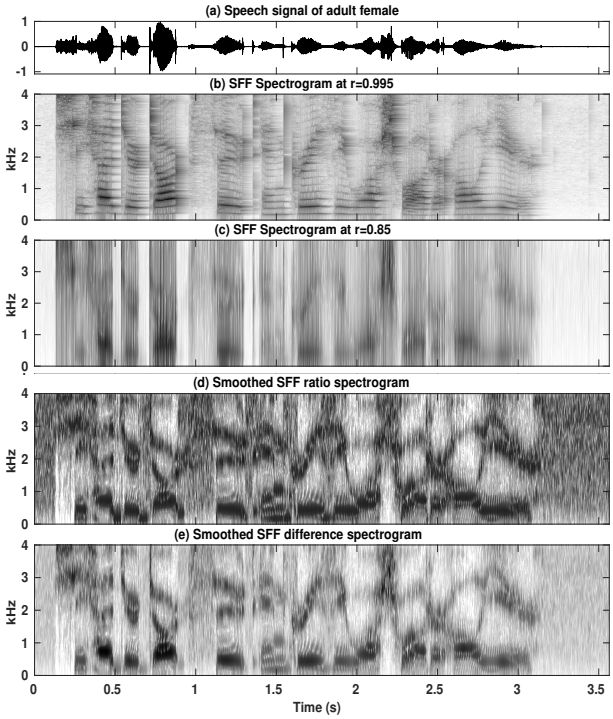

Figure 2: *Illustration of formant enhancement for adult female voice. (a) Speech signal. (b) Spectrogram at $r = 0.995$. (c) Spectrogram at $r = 0.85$. (d) Smoothed SFF ratio spectrogram. (e) Smoothed SFF difference spectrogram.*

$$e_{kr}[n] = \frac{e_{k1}[n]}{e_{k2}[n]}, \tag{6}$$

for all k at each n. Note that the value of $\Delta r = 0.0001$ is used here, although, although any value in the range 0.001 to 0.00001 can be chosen. The modified SFF spectrogram $v_{kr}[n]$ is obtained as follows.

$$\begin{aligned} v_{kr}[n] &= e_{kr}[n] - 1, \; if \; e_{kr}[n] > 1, \\ &= 0, \qquad\quad if \; e_{kr}[n] < 1. \end{aligned} \tag{7}$$

The plot of $v_{kr}[n]$ for all k and n gives the SFF ratio spectrogram.

The spectral resolution of a wideband SFF spectrogram can also be improved by subtracting the SFF spectrograms at r and $r - \Delta r$. The SFF difference spectrum $e_{ks}[n]$ is given by

$$e_{ks}[n] = e_{k1}[n] - e_{k2}[n], \tag{8}$$

for all k at each n. The modified SFF difference spectrum is given by

$$\begin{aligned} v_{ks}[n] &= e_{ks}[n], \; if \; e_{ks}[n] > 0, \\ &= 0, \qquad if \; e_{ks}[n] < 0. \end{aligned} \tag{9}$$

The plot of $v_{ks}[n]$ for all k at each n gives the SFF difference spectrogram.

To highlight the formant information further, the modified SFF ratio spectrogram or the modified SFF difference spectrogram is smoothed by taking the mean of these spectrogram values over M consecutive time instants. Fig. 2(d) and Fig. 2(e) show the smoothed SFF ratio and smoothed SFF difference spectrograms, respectively, for $M = 8$, corresponding to averaging over 1 ms at 8 kHz sampling rate.

The formant information can be seen clearly in the expanded part of a segment of Figs. 2(d) and 2(e) in Figs. 3(b) and 3(c), respectively. The figures show that the formant information is better displayed in the smoothed SFF ratio (Fig. 2(d))

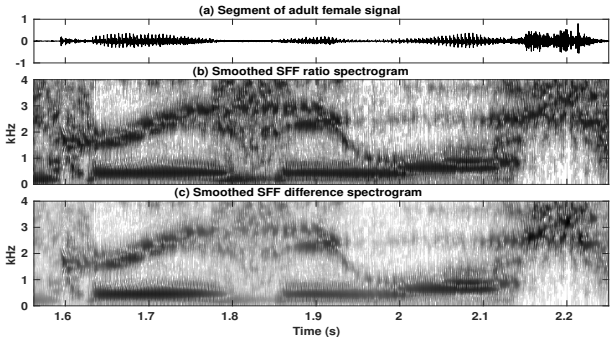

Figure 3: *Illustration of formant enhancement for a segment of adult female voice. (a) Speech signal. (b) Smoothed SFF ratio spectrogram. (c) Smoothed SFF difference spectrogram.*

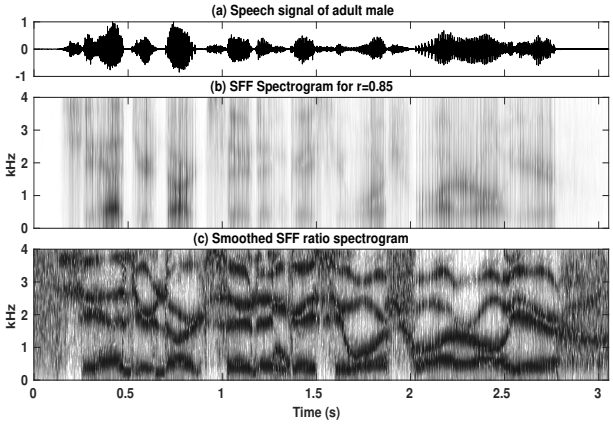

Figure 4: *Illustration of formant enhancement for adult male voice. (a) Speech signal. (b) SFF spectrogram for* $r = 0.85$*. (c) Smoothed SFF ratio spectrogram.*

and SFF difference (Fig. 2(e)) spectrograms, than in the SFF spectrograms shown in Fig. 2(b) and 2(c). Since both SFF ratio and SFF difference spectrograms give similar results, we use the SFF ratio spectrogram for the illustrations given in the next section.

4. Results for different speech utterances

The effectiveness of the smoothed SFF ratio spectrogram for enhancing the formant information is examined for a male voice (normal pitch) and for children voices (high pitch). The result for a male voice is shown in Fig. 4. In the figure one can clearly observe that the formants are well resolved throughout the utterance, without the effects of harmonics as in the narrowband spectrogram, and also without the effects of poor resolution as in the wideband spectrogram. One can also observe the rapid transitions in the formant contours. Notice that in a few cases the continuity of formants at high frequency can be observed in the fricative regions. For nasal sounds, the lowering of the first formant can be seen. The voice bar in the case of voiced stops cannot be observed, as they are eliminated by the choice of low value of r in the SFF analysis. The SFF spectrogram for low value of r will not capture the periodicities of the voice bar. It is important to note that for male voices with normal pitch, the STFT wideband and narrowband spectrograms also show the formant information clearly. Hence the proposed method may not have any specific advantage in some of those cases. It is interesting to note that enhancement of the formant information takes place for high pitched (200-400 Hz) voices like female and children voices. This is because we use a very low value of r to eliminate the effects of harmonics due to small pitch periods, but enhance the formants using the proposed SFF ratio spectrogram. Fig. 5 and Fig. 6 show the formants and formant transitions in the display of the SFF ratio spectrograms, compared to the wideband equivalent SFF spectrograms for two high pitched voices of a female child and a male child, both about 4 years of age.

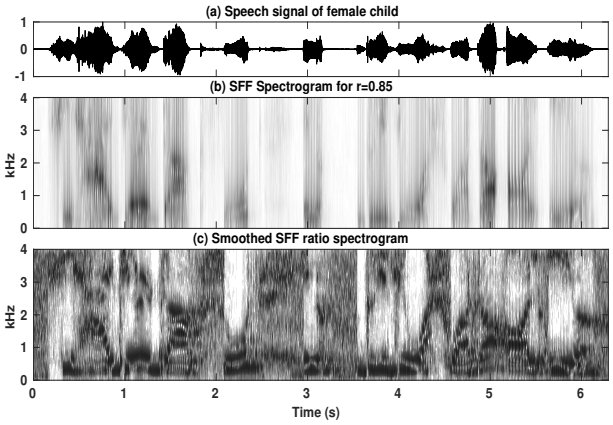

Figure 5: *Illustration of formant enhancement for female child voice. (a) Input signal. (b) SFF spectrogram for* $r = 0.85$*. (c) Smoothed SFF ratio spectrogram.*

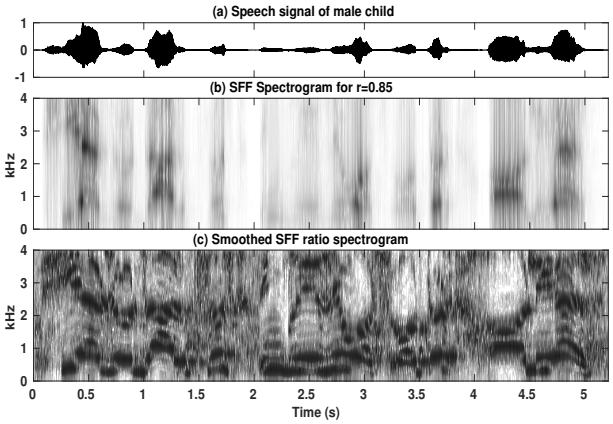

Figure 6: *Illustration of formant enhancement for male child voice. (a) Input signal. (b) SFF spectrogram for* $r = 0.85$*. (c) Smoothed SFF ratio spectrogram.*

5. Summary and conclusions

Extraction of formant information from speech is normally based on the assumptions on the locations of peaks in the spectral envelope and continuity constraints for tracking. Wideband and narrowband spectrograms have limitation of masking formant information due to poor frequency resolution and due to the effects of harmonics, respectively. We proposed an SFF-based method to derive a spectrogram which allows us to observe the formant information with good resolution in both frequency and time domains. The method improves the frequency resolution of the wideband equivalent of the SFF spectrogram. The method involves computing the ratio or difference of the SFF spectra computed at two closely spaced values of the parameter r used in the SFF analysis. The formant information is clearly visible in the SFF ratio and difference spectrograms, for different types of voices. Rapid formant transitions and formant breaks are clearly visible in the SFF ratio and difference spectrograms, especially in the high-pitched voices. The challenge of extracting the formants and formant contours still remains, as it is difficult to filter out the information not related to formants from these SFF ratio or difference spectrograms.

6. References

[1] M. Lee, J. P. H. van Santen, B. Möbius, and J. P. Olive, "Formant tracking using context-dependent phonemic information," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 5-2, pp. 741–750, 2005.

[2] P. Boersma and D. Weenink, "Praat, a system for doing phonetics by computer," vol. 5, 2002, pp. 341–345.

[3] B. Munson and N. Solomon, "The effect of phonological neighborhood density on vowel articulation," *Journal of speech, language, and hearing research : JSLHR*, vol. 47, pp. 1048–58, 2004.

[4] M. Narendranath, H. A. Murthy, S. Rajendran, and B. Yegnanarayana, "Transformation of formants for voice conversion using artificial neural networks," *Speech Communication*, vol. 16, no. 2, pp. 207–216, 1995.

[5] l. Deng and J. Ma, "Spotaneous speech recognition using a statistical coarticulatory model for the vocal-tract-resonance dynamics," *The Journal of the Acoustical Society of America*, vol. 108, pp. 3036–48, 2001.

[6] L. Deng, X. Cui, R. Pruvenok, J. Huang, S. Momen, Y. Chen, and A. Alwan, "A database of vocal tract resonance trajectories for research in speech processing," in *International Conference on Acoustics Speech and Signal Processing Proceedings*, vol. 1. IEEE, 2006, pp. I–I.

[7] J. Makhoul, "Linear prediction: A tutorial review," *Proceedings of the IEEE*, vol. 63, no. 4, pp. 561–580, 1975.

[8] J. M. Anand, S. Guruprasad, and B. Yegnanarayana, "Extracting formants from short segments of speech using group delay functions," in *INTERSPEECH*, 2006, pp. 1009–1012.

[9] H. A. Murthy and B. Yegnanarayana, "Formant extraction from group delay function," *Speech Communication*, vol. 10, no. 3, pp. 209–221, 1991.

[10] B. Yegnanarayana, "Formant extraction from linear-prediction phase spectra," *The Journal of the Acoustical Society of America*, vol. 63, no. 5, pp. 1638–1640, 1978.

[11] L. Deng and C. D. Geisler, "A composite auditory model for processing speech sounds," *The Journal of the Acoustical Society of America*, vol. 82, no. 6, pp. 2001–2012, 1987.

[12] K. Xia and C. Y. Espy-Wilson, "A new strategy of formant tracking based on dynamic programming," in *INTERSPEECH*, 2000.

[13] G. Kopec, "Formant tracking using hidden Markov models and vector quantization," *IEEE Transactions on Acoustics, Speech, and Signal processing*, vol. 34, no. 4, pp. 709–729, 1986.

[14] D. D. Mehta, D. Rudoy, and P. J. Wolfe, "Kalman-based autoregressive moving average modeling and inference for formant and antiformant tracking," *The Journal of the Acoustical Society of America*, vol. 132, no. 3, pp. 1732–1746, 2012.

[15] L. Deng, L. J. Lee, H. Attias, and A. Acero, "Adaptive Kalman filtering and smoothing for tracking vocal tract resonances using a continuous-valued hidden dynamic model," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 1, pp. 13–23, 2006.

[16] Y. Dissen, J. Goldberger, and J. Keshet, "Formant estimation and tracking: A deep learning approach," *The Journal of the Acoustical Society of America*, vol. 145, no. 2, pp. 642–653, 2019.

[17] D. OShaughnessy, "Formant estimation and tracking," in *Springer handbook of speech processing*. Springer, 2008, pp. 213–228.

[18] Y. Bayya and D. N. Gowda, "Spectro-temporal analysis of speech signals using zero-time windowing and group delay function," *Speech Communication*, vol. 55, no. 6, pp. 782–795, 2013.

[19] R. Prasad and B. Yegnanarayana, "Determination of glottal open regions by exploiting changes in the vocal tract system characteristics," *The Journal of the Acoustical Society of America*, vol. 140, pp. 666–677, 07 2016.

[20] G. Aneeja and B. Yegnanarayana, "Single frequency filtering approach for discriminating speech and nonspeech," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 4, pp. 705–717, 2015.