IMPROVEMENT OF PROBABILISTIC ACOUSTIC TUBE MODEL FOR SPEECH DECOMPOSITION

Yang Zhang[1], *Zhijian Ou*[2], *Mark Hasegawa-Johnson*[1]

1. University of Illinois, Urbana-Champaign, Department of Electrical and Computer Engineering
2. Tsinghua University, Department of Electronic Engineering

yzhan143@illinois.edu, ozj@tsinghua.edu.cn, jhasegaw@illinois.edu

ABSTRACT

Current model-based speech analysis tends to be incomplete - only a part of parameters of interest (e.g. only the pitch or vocal tract) are modeled, while the rest that might as well be important are disregarded. The drawback is that without joint modeling of parameters that are correlated, the analysis on speech parameters may be inaccurate or even incorrect. Under this motivation, we have proposed such a model called PAT (Probabilistic Acoustic Tube), where pitch, vocal tract and energy are jointly modeled. This paper proposes an improved version of PAT model, named PAT2, where both signal and probabilistic modeling are tremendously renovated. Compared to related works, PAT2 is much more comprehensive, which incorporates mixed excitation, glottal wave and phase modeling. Experimental results show its ability in decomposing speech into desirable parameters and its potential for speech synthesis.

***Index Terms*—** Probabilistic generative model, model-based speech processing, speech modeling

1. INTRODUCTION

Most speech processing tasks (e.g. pitch estimation, speech recognition, source separation and so on) require a probabilistic model of speech. However, current model-based speech analysis tend to be incomplete - they tend to model only a part of parameters of interest, and disregard others that might also be important.

The drawback is that without jointly modeling parameters that are correlated, the analysis on speech parameters may be inaccurate or even incorrect. For example, Kameoka [1] noted that pitch and spectral envelope have a "chicken and egg" relationship and should be estimated jointly. Stephenson [2] pointed out that cepstral-based features are sensitive to "auxiliary information" such as pitch and energy. An extreme example would be that current vocal tract estimation, such as LPC and MFCC, is always corrupted by 'spectral tilt' induced by glottal wave, and can be correctly estimated only when both are jointly modeled.

Under this motivation, we have proposed a model called PAT (Probabilistic Acoustic Tube)[3], where pitch, vocal tract and energy are jointly modeled. Preliminary experiments show that PAT has a good potential for various speech processing tasks, such as pitch tracking, speech enhancement etc.

There is, however, room for improvement. This paper proposes an improved version of the PAT model, named PAT2, where both signal and probabilistic modeling are tremendously renovated. Specifically, there are several highlights. *First*, the model incorporates breathiness and glottal vibration, based on recent findings in the study of speech production [4]. *Second*, instead of modeling the magnitude spectrum only, PAT2 incorporates phase modeling and so completely defines a probabilistic model for the complex spectrum of speech. *Third*, instead of setting different models for voiced and unvoiced speech, as in many speech processing methods, PAT2 makes U/V states a continuum by introducing voiced amplitude and unvoiced amplitude, which is closer to the nature of speech.

The rest of the paper is organized as follows. Section 2 and 3 describe signal modeling and probabilistic modeling of PAT2 respectively. Section 4 gives experimental results which demonstrate the effectiveness of PAT2. Finally, in section 5 we discuss related work and point out future work.

Notations: We use lower-case letter with bracketed index n, e.g. $x[n]$, to denote time domain discrete signals; upper case letter with parenthesized index ω, e.g. $X(\omega)$, to denote its DTFT; bold lower-case letter, e.g. $\boldsymbol{x}$, for *column* vectors and bold upper-case letter, e.g. $\boldsymbol{X}$, for matrices; IDTFT$[\cdot]$ for inverse DTFT operator; $\circledast$ for circular convolution.

2. SIGNAL MODELING OF PAT2

2.1. The Source-Filter Model with Mixed Excitation

PAT2, essentially, is a source-filter model. Yet unlike the common source-filter model, which switches between the voiced excitation (glottal vibration) and unvoiced excitation (breath), PAT2 allows mixed excitations. This is because even for voiced speech, there is a significant amount breathiness [5]. The unvoiced case in PAT2 is thus reduced to a special case of voiced speech where the amplitude of voiced excitation drops to zero.

Formally, suppose rectangular window is chosen for each speech frame, the DTFT of speech, $S_t(\omega)$, can be represented as

$$S_t(\omega) = [a_t V_t(\omega) + b_t U_t(\omega)]\, H_t(\omega) \circledast W_t(\omega) \tag{1}$$

where t is the frame index; $V_t(\omega)$ is the DTFT of voiced excitation; $U_t(\omega)$ is the DTFT of breath noise; $H_t(\omega)$ is the vocal tract transfer function; $W_t(\omega)$ is the DTFT of rectangular window; a_t and b_t are the voiced amplitude and unvoiced amplitude respectively. Rectangular window may seem an uncommon choice in speech processing due to its high sidelobes, but in probabilistic modeling, all windows are equivalent as long as the distribution is determined correctly. As will be shown in the next section, rectangular window leads to the simplest form of distribution.

The voiced excitation can be modeled as periodic repetition of glottal wave, or say, glottal wave convolved with a pulse train. In the

This work is supported by the Illinois Innovation Initiative and NSFC under grant No. 61075020.

frequency domain, it is represented as

$$V_t(\omega) = G_t(\omega) e^{-j\omega\tau_t} \sum_k \delta(\omega - k\omega_{0t}) \tag{2}$$

where $G_t(\omega)$ is the DTFT of glottal wave of one cycle; $\delta(\omega)$ is the dirac delta function; ω_{0t} is the fundamental frequency in radians; τ_t, is the group delay, namely the time relative to the beginning of the frame when the first pulse appears.

The breath noise is simply white Gaussian noise with unit variance in the time domain:

$$u_t[n] = \text{IDTFT}[U_t(\omega)] \overset{iid}{\sim} \mathcal{N}(0, 1) \tag{3}$$

whose statistical behavior in the frequency domain will be discussed in section 3.

With this framework, signal modeling of PAT2 is reduced to modeling of $G_t(\omega)$ and $H_t(\omega)$, and they will be discussed in detail in the following subsections respectively.

2.2. All-pole Model For Glottal Wave

We adopt the common practice that incorporates radiation effect of speech by taking first-order finite difference of glottal wave [6]. Glottal derivative contains an open phase, a return phase and a closed phase. The connection between open phase and return phase, where short time energy of speech is usually greatest, is called GCI (glottal closure instant) [4].

It has been shown in [7] that coarse structure of glottal flow can be approximated by three poles: a pair of conjugate maximum-phase poles (outside unit circle) and a real minimum-phase pole (inside unit circle), which has a very close link to the famous LF model [8]. If we assume an impulse input at GCI, the maximum-phase pole pairs model the open phase, and the minimum-phase pole models the return phase.

We apply the three-pole model to PAT2, namely

$$G_t(\omega) = \frac{1}{(1 + 2g_{1t}\cos\beta_t e^{-j\omega} + g_{1t}^2 e^{-2j\omega})(1 + g_{2t}e^{-j\omega})} \tag{4}$$

where g_{1t}, β_t and g_{2t}, are the magnitude and phase of the maximum-phase pole pair, and the magnitude of the minimum-phase real pole.

2.3. Mel-Frequency Complex Cepstral Coefficient (MFC^3)

MFCC is widely used by speech recognition systems to represent the *magnitude* of vocal tract transfer function. However, *complex* vocal tract transfer function is modeled in PAT2, and we obtain mel-frequency *complex* cepstral coefficient, abbreviated as MFC^3.

MFC^3 is extracted from the mel-frequency complex cepstrum of $H_t(\omega)$, defined as

$$\hat{h}_t[\hat{n}] = \text{IDTFT}[\log(H_t(\tilde{\omega}))] \tag{5}$$

where $\hat{n}$ is quefrency; $\tilde{\omega}$ is the mel-frequency:

$$\tilde{\omega} = m(\omega) = \begin{cases} \omega & \omega < 2000\pi \\ \log\left(\frac{\omega}{1400\pi} + 1\right) \times 2254\pi & \text{otherwise} \end{cases} \tag{6}$$

According to previous study [9], vocal tract can be well modeled by a minimum phase system. Thus, it can be proved [10] that $\hat{h}_t[\hat{n}]$ is right-sided, namely 0 when $\hat{n} < 0$, and if group delay and the sign of gain of $H(\omega)$ are properly removed, $\hat{h}_t[\hat{n}]$ decays at the rate of $1/\hat{n}$. According to (1) and (2), the sign of gain is controlled by a_t and group delay by τ_t. Therefore, we can use $\hat{h}_t[\hat{n}], 0 < \hat{n} \leq K$ for small K, named MFC^3, to represent $H_t(\omega)$:

$$H_t(\omega) = \exp\left(\sum_{\hat{n}=1}^{K} \hat{h}_t[\hat{n}] \exp(-jm(\omega)\hat{n})\right) \tag{7}$$

where $K = 26$ in our experiment. The 0-th coefficient is removed because amplitude is already modeled by a_t and b_t.

So far, the signal model has been completely established by (1), (2), (3), (4) and (7). which is the basis of the probabilistic model introduced in the next section. The parameter set Θ is

$$\Theta = \bigcup_t \left(\{a_t, b_t, \omega_{0t}, \tau_t, g_{1t}, \beta_t, g_{2t}\} \cup \bigcup_{\hat{n}} \left\{\hat{h}_t[\hat{n}]\right\}\right) \tag{8}$$

3. PROBABILISTIC MODELING OF PAT2

3.1. Compact Real DFT Representation

We will switch to DFT representation from DTFT. DFT of a real speech signal is conjugate symmetric, and thus we only need to use the first half of the DFT coefficients. Formally, denote N as frame length. If N is even, define

$$\begin{aligned} \boldsymbol{s}_t = \Bigg[& S_t^{(r)}\left(\frac{2\pi 0}{N}\right), ..., S_t^{(r)}\left(\frac{2\pi(N/2)}{N}\right), \\ & S_t^{(i)}\left(\frac{2\pi}{N}\right), ..., S_t^{(i)}\left(\frac{2\pi(N/2-1)}{N}\right)\Bigg]^T \end{aligned} \tag{9}$$

where superscript (r) and (i) denotes real part and imaginary part respectively. $S_t^{(i)}(0)$ and $S_t^{(i)}(\pi)$ are not included because they are constantly 0. This length N vector contains exactly the same information as the time domain signal. We call it the compact real DFT representation of $S_t(\omega)$.

3.2. Likelihood of Speech Complex Spectrum

Considering that there are unmodelled speech effects, such as jitter and shimmer, and these effects tend to concentrate on high frequency, we switch to modeling mel-scale representation of speech $\boldsymbol{s}_t$ to minimize the error. According to (1), we have

$$\tilde{\boldsymbol{s}}_t \equiv \boldsymbol{F}\boldsymbol{s}_t = a_t\boldsymbol{F}\boldsymbol{\xi}_t + b_t\boldsymbol{F}\boldsymbol{\eta}_t \tag{10}$$

where $\boldsymbol{F}$ is a matrix containing rows of overlapping triangular windows, whose end points are uniform in mel-scale; $\boldsymbol{\xi}_t$ and $\boldsymbol{\eta}_t$ are compact real DFT representation of $V_t(\omega) H_t(\omega) \circledast W_t(\omega)$ and $U_t(\omega) H_t(\omega) \circledast W_t(\omega)$ respectively.

$\boldsymbol{\eta}_t$ is the only random variable, whose distribution will now be derived. First, the rectangular window can be removed because it has no impact on DFT. Therefore, $\boldsymbol{\eta}_t$ can be further represented as

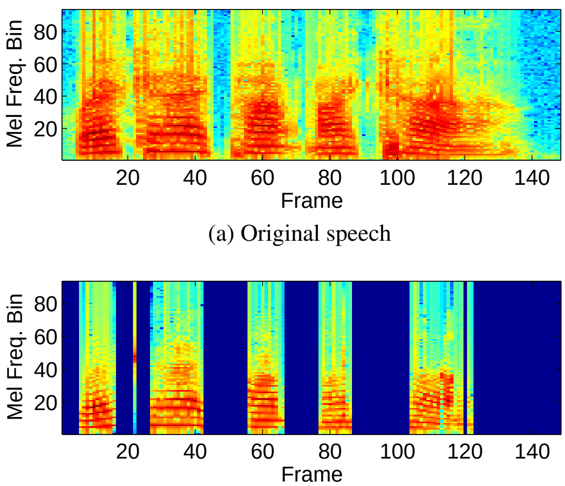
$$\boldsymbol{\eta}_t = \boldsymbol{H}_t\boldsymbol{u}_t \tag{11}$$

where $\boldsymbol{u}_t$ is the compact real DFT representation of $U_t(\omega)$ and $\boldsymbol{H}_t$ is the 4-block (2 by 2) matrix that achieves complex multiplication between $U_t(\omega)$ and $H_t(\omega)$.

DFT is an orthogonal transform. With some simple manipulations and (3), it can be proved that $\boldsymbol{u}_t$ is a standard multivariate Gaussian variable, i.e. with zero mean and identity covariance matrix. Therefore

$$\tilde{\boldsymbol{s}}_t \sim \mathcal{N}\left(a_t\boldsymbol{F}\boldsymbol{\xi}_t, b_t^2\boldsymbol{F}\boldsymbol{H}_t\boldsymbol{H}_t^T\boldsymbol{F}^T|\Theta\right) \tag{12}$$

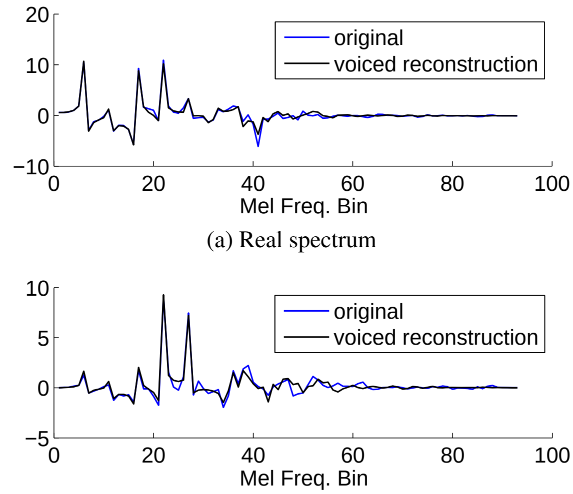
which defines the likelihood of the observed speech frame.



(a) Original speech

(b) Reconstructed voiced speech; dark blue represents $-\infty$

Fig. 1. Comparison of log magnitude spectrogram of reconstructed voiced speech and original speech.



(a) Real spectrum

(b) Imaginary spectrum

Fig. 2. Comparison of real and imaginary spectrum for frame 35.

3.3. Parameter Priors

To ensure parameters vary smoothly across frames, it is useful to set smoothing priors for them. We assume conditional Gaussian priors

$$\log P_{\theta_t|\theta_{t-1}}(u|v) \propto -\frac{(u-v)^2}{\sigma_\theta^2} \tag{13}$$

for all parameters $\theta_t \in \Theta$ except for a_t and τ_t, where σ_θ^2 are, for now, hand-tuned hyper-parameters[1].

The voiced amplitude of frame t, a_t, is crucial for U/V decision, because by assumption, $a_t = 0$ corresponds to unvoiced speech. In actuality, however, the estimate of a_t is rarely exactly equal to 0 due to corruption of noise. We set a "bonus" in prior for a_t being equal 0:

$$\begin{aligned}\log P_{a_t|a_{t-1}}(u|v) \propto -\frac{(u-v)^2}{\sigma_a^2} + B\mathbb{1}[u=0] \\ -C\mathbb{1}[\mathbb{1}[u=0] \neq \mathbb{1}[v=0]]\end{aligned} \tag{14}$$

where $\mathbb{1}[\cdot]$ is indicator function. The third term imposes extra cost for constantly jumping between U/V states. B and C are hand-tuned parameters[2]. Under MAP (maximum a posteriori) scheme, this prior should attract small a_t's to strict 0.

For τ_t of all frames and other speech parameters of the 0-th frame, we assume noninformative priors. (12), (13) and (14) define the probabilistic model for PAT2. Parameters are estimated with MAP using gradient ascent method.

4. EXPERIMENTS

In this section, a set of preliminary results are displayed to demonstrate versatility and representation power of PAT2. These results are obtained by running PAT2 on an utterance "Where can I park my car" by a male speaker in the Edinburgh speech corpus [11]. The log spectrogram is displayed in fig.1(a). We will see that PAT2 does well in speech modeling (especially for phase) and decomposing the speech into desirable parameters.

4.1. Speech Reconstruction

As an illustration of general performance of PAT2, speech reconstruction is performed, where all the speech parameter estimates are assembled to reconstruct the *voiced* part of speech $a_t \boldsymbol{F}\boldsymbol{\xi}_t$ as in equation (10). If the reconstructed speech is close to the original speech in voiced segments, we can say PAT2 generally models speech well.

Fig.1(b) displays the log spectrogram of voiced reconstruction. As can be seen, there are segments where the voiced reconstruction is strictly zero. These segments correspond to unvoiced and silent segments judged by PAT2, which demonstrates PAT2's U/V decision. In voiced segments, pitch harmonics and formant structure, especially in low frequencies, are very close to those in original speech. In mid and high frequencies, voiced reconstruction has lower energy than the original speech. This is because PAT2 regards the excluded energy as unvoiced, or breath excited; according to [5], breath energy is likely to dominate mid and high frequencies.

To have a clearer view, reconstruction of frame 35 is compared with the original speech in fig.2. Both real spectrum and imaginary spectrum are compared, rather than only considering the magnitudes. We can see that the reconstruction almost overlaps with the original in low frequencies in both spectra, which shows that PAT2 models *phase* of speech very accurately. This result demonstrates PAT2's potential for speech synthesis and model-based separation.

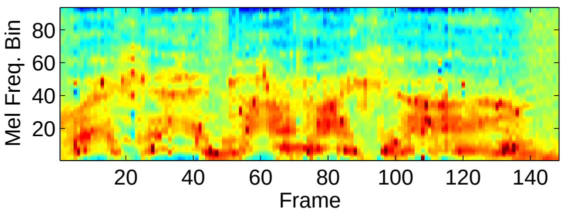
4.2. 'Glottal Free' Vocal Tract Transfer Function

This result illustrates how PAT2 disentangles parameters. As discussed in section 1, MFCC essentially mixes glottal wave and vocal tract transfer function, and their separation cannot be obtained without a unified model like PAT2. Therefore, PAT2 provides some insights of disentangled vocal tract.

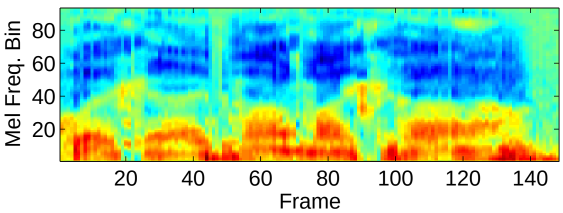
Fig.3 compares the vocal tract transfer function estimated by PAT2 with the spectral envelope estimated by MFCC. An immediate

[1] In this experiment, we set $\sigma_{\omega 0}^2 = 1$, $\sigma_a^2 = \sigma_b^2 = 0.01$, $\sigma_{\hat{h}}^2 = \sigma_{g1}^2 = \sigma_{g2}^2 = \sigma_\beta^2 = 0.1$, $\sigma_\tau^2 = 1e-5$.

[2] In this experiment, we set $B = 50$, $C = 10$.

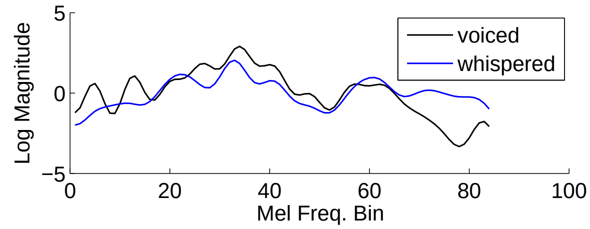


(a) Log magnitude of vocal tract transfer function estimated by PAT2

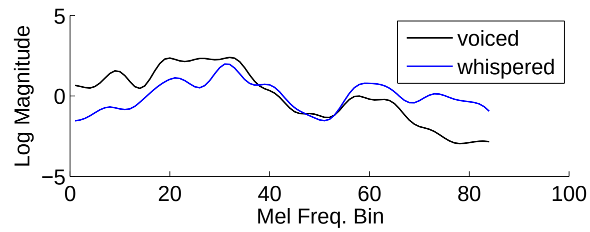


(b) Log magnitude of spectral envelope estimated by MFCC

Fig. 3. Comparison of spectral envelope representation of PAT2 and MFCC.



(a) PAT2 estimation



(b) MFCC estimation

Fig. 4. The means of the estimated vocal tract frequency response / spectral envelope for a voiced-excited and a whispered utterance of /ah/

observation is that the spectral tilt induced by glottal wave is largely removed in PAT2's representation. Notice that the removal is not heuristic, but based on phase information (maximum phase component). MFC[3] thus has the potential to be a better feature for speech recognition because glottal variation and breathiness affect spectral envelope, but do not change the vocal tract.

To further illustrate this point, we perform another experiment where 2 extreme utterances of /ah/ are recorded, one uttered with voiced excitation and the other whispered. The idea is that the vocal tract shapes in both cases are similar, but one has spectral tilt and the other doesn't. It is expected that PAT2 model would give more consistent estimates of vocal tract of the two cases than MFCC does.

Fig.4 compares the mean of the envelope estimates of both cases by the two methods. It turns out that both MFCC and PAT2 have almost the same envelope estimates for the whispered case, but very different for voiced. PAT2 has much more consistent estimates for both cases, especially in the mid frequency. The norm of the differences between the means of the estimates for the two cases is 10.93 for PAT2, as opposed to 13.15 for MFCC.

4.3. GCI Location

GCI estimation is a good indication of PAT2's ability to model phase and pitch tracking. According to section 2, τ_t is the delay of the first GCI relative to the beginning of frame t. Also we know that GCI are periodic at the fundamental frequency. Then, estimated GCI locations of frame t are thus $\tau_t + 2k\pi/\omega_{0t}$, where k is a nonnegative integer. Since GCI's of different frames are estimated separately, we can judge the accuracy by checking: 1) if GCI's of different frames are consistent, i.e. if they form a quasi-periodic sequence; 2) if they appear at energy bursts in the original speech.

Fig.5 plots GCI locations as impulses and the original speech waveform. As can be seen, GCI's, around 3 or 4 instances in each frame, form a periodic signal with rare exceptions. What's more, they tend to appear consistently at the largest negative to positive jump within a period in the original speech wave, where short-time energy is generally greatest. This result shows that PAT2 can control for group delay and pitch well, and thus achieves similar performance to pitch-synchronous analysis.

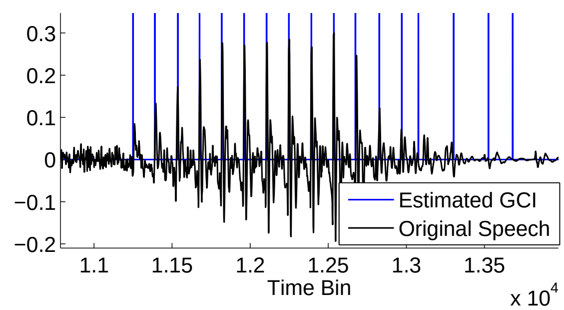


Fig. 5. Estimation of GCI location of the utterance 'park'.

5. RELATED WORK AND CONCLUSION

Previous works on speech generative models include [12][13]. There are other attempts to jointly model speech parameters, such as the STRAIGHT model [14] and the compound model [1]. Compared to these previous models, PAT2 is probabilistic and much more comprehensive, which incorporates mixed excitation, glottal wave and phase modeling.

In conclusion, we proposed a comprehensive generative model for speech, and showed its ability in decomposing the speech into desirable parameters and its potential for speech synthesis. However, there is still room for further improvement, the greatest being glottal wave modeling. Although LF model is good for coarse structure of glottal wave, it does not model fine structure, which may introduce some errors in PAT2. Also, the priors of speech parameters are set heuristically without a standard training procedure. Nonetheless, we believe that with improvement, PAT2 will find its way into many speech processing tasks and speech research applications.

6. REFERENCES

[1] Hirokazu Kameoka, Nobutaka Ono, and Shigeki Sagayama, "Speech spectrum modeling for joint estimation of spectral envelope and fundamental frequency," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 18, no. 6, pp. 1507–1516, 2010.

[2] Todd A Stephenson, Mathew Magimai Doss, and Hervé Bourlard, "Speech recognition with auxiliary information," *Speech and Audio Processing, IEEE Transactions on*, vol. 12, no. 3, pp. 189–203, 2004.

[3] Zhijian Ou and Yang Zhang, "Probabilistic acoustic tube: a probabilistic generative model of speech for speech analysis/synthesis," in *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2012, pp. 841–849.

[4] Thomas F Quatieri, *Discrete-time speech signal processing*, Pearson Education India, 2002.

[5] Dennis H Klatt and Laura C Klatt, "Analysis, synthesis, and perception of voice quality variations among female and male talkers," *the Journal of the Acoustical Society of America*, vol. 87, pp. 820, 1990.

[6] Thomas Drugman, Baris Bozkurt, and Thierry Dutoit, "Causal–anticausal decomposition of speech using complex cepstrum for glottal source estimation," *Speech Communication*, vol. 53, no. 6, pp. 855–866, 2011.

[7] William R Gardner and Bhaskar D Rao, "Noncausal all-pole modeling of voiced speech," *Speech and Audio Processing, IEEE Transactions on*, vol. 5, no. 1, pp. 1–10, 1997.

[8] Gunnar Fant, Johan Liljencrants, and Qi-guang Lin, "A four-parameter model of glottal flow," *STL-QPSR*, vol. 4, no. 1985, pp. 1–13, 1985.

[9] Osamu Fujimura, "Analysis of nasal consonants," *The Journal of the Acoustical Society of America*, vol. 34, pp. 1865, 1962.

[10] Alan V Oppenheim, Ronald W Schafer, John R Buck, et al., *Discrete-time signal processing*, vol. 5, Prentice Hall Upper Saddle River, 1999.

[11] Paul C Bagshaw, Steven M Hiller, and Mervyn A Jack, "Enhanced pitch tracking and the processing of f0 contours for computer aided intonation teaching.," in *Proc. Eurospeech*. International Speech Communication Association, 1993.

[12] Kannan Achan, Sam Roweis, Aaron Hertzmann, and Brendan Frey, "A segment based probabilistic generative model of speech," in *Acoustics, Speech, and Signal Processing, 2005. Proceedings.(ICASSP 2005). IEEE International Conference on*. IEEE, 2005, vol. 5, pp. v–221.

[13] Francis R Bach and Michael I Jordan, "Discriminative training of hidden markov models for multiple pitch tracking [speech processing examples]," in *Acoustics, Speech, and Signal Processing, 2005. Proceedings.(ICASSP 2005). IEEE International Conference on*. IEEE, 2005, vol. 5, pp. v–489.

[14] Hideki Kawahara, Masanori Morise, Toru Takahashi, Ryuichi Nisimura, Toshio Irino, and Hideki Banno, "Tandem-straight: A temporally stable power spectral representation for periodic signals and applications to interference-free spectrum, f0, and aperiodicity estimation," in *Acoustics, Speech and Signal Processing, 2008.(ICASSP 2008). IEEE International Conference on*. IEEE, 2008, pp. 3933–3936.