INCORPORATING AM-FM EFFECT IN VOICED SPEECH FOR PROBABILISTIC ACOUSTIC TUBE MODEL

Yang Zhang† *Zhijian Ou*⋆ *Mark Hasegawa-Johnson*†

† University of Illinois, Urbana-Champaign, Department of Electrical and Computer Engineering
⋆ Tsinghua University, Department of Electronic Engineering
yzhan143@illinois.edu, ozj@tsinghua.edu.cn, jhasegaw@illinois.edu

ABSTRACT

A complete speech model can improve performance for many speech applications. Probabilistic Acoustic Tube (PAT) is a probabilistic generative model of speech that has been shown potentially useful in a number of speech processing tasks. A point missing in previous PAT models is that they overlook AM/FM effect in voiced speech, which is in fact common and non-negligible. In this paper, we significantly improve the voiced modeling of PAT with a probabilistic model of AM/FM effect, which is developed from Bayesian Spectrum Estimation method. Experiments show that the new PAT is able to fit the voiced speech spectrum with greater accuracy in the presence of AM/FM effect.

Index Terms— Speech modeling, speech analysis, AM/FM, generative model

1. INTRODUCTION

Most speech processing tasks benefit from a complete model of speech that fully takes into account important speech elements instead of partial modeling and feature extraction. For example, in speech analysis, it is shown in [1] that estimation of pitch and spectral envelope should be performed jointly. In speech synthesis, it is shown that jointly modeling glottal source and vocal tract improves the quality of parametric speech synthesis [2]. In source separation, a complete speech model can more accurately define the sample space of clean signal and therefore is better able to recover clean speech [3, 4].

As a result, there have been many efforts on building complete speech models. The STRAIGHT model [5] jointly models pitch, glottal source and spectral envelope, which is proven effective in speech modification and resynthesis. Degottex et. al. [6] proposed a speech model with mixed excitation and adapted vocal tract estimate, which can be used for speech resynthesis, breathiness modification and pitch adjustment.

In most studies, although different speech elements are considered within a signal model, their estimations are still conducted separately. This may result in inconsistencies between analysis and synthesis. Also, few of these studies obtained a unified probabilistic model of speech. In contrast, we proposed a complete generative model of speech in [7, 8], named Probabilistic Acoustic Tube (PAT) model, which jointly models breathiness, glottal excitation and vocal tract in a probabilistic modeling framework, and notably with phase information. Preliminary experiments have demonstrated good potential of PAT in a number of speech applications.

This project is supported by AHRQ grant R21-HS022948, and NSFC grant 61473168.

A remarkable point missing in previous PAT models is that voiced speech is assumed to be perfectly stationary, i.e. it is a strictly periodic signal, while in fact variations within a single voiced speech frame are common and non-negligible [9]. Two main variations are pitch jitter and amplitude shimmer, referring to the phenomena that the pitch period may randomly vary and the amplitude of the airflow velocity within a glottal cycle may differ across successive periods in voiced speech, due, perhaps, to time-varying characteristics of vocal folds. Jitter and shimmer give voiced speech its naturalness, but introduce AM/FM effect. AM/FM widens the harmonic pulses in voiced speech spectrum, and the widening becomes more significant as the frequency increases. A failure to account for this AM/FM effect results in over-estimation of environmental and aspiration noise in previous PAT models, especially in frequency band above 2kHz.

Therefore, in this paper, we significantly improve voiced modeling of PAT by introducing a probabilistic model on AM/FM effect, which is an adaptation of traditional Bayesian Spectrum Estimation (BSE) [10]. Experiments show that the new PAT, called PAT3, is able to successfully fit those portions of voiced spectrum that are caused by AM/FM effect and mistakenly ascribed to background or aspiration noise by previous PAT models.

The rest of the paper is organized as follows. Section 2 gives a brief on the signal and probabilistic models of PAT3; section 3 derives the AM/FM model for PAT3; section 4 demonstrates PAT3's ability in modeling speech with AM/FM effect; section 5 concludes the paper and points out future directions.

Notations: Lower-cased letters with bracketed t, $a[t]$, denote discrete time domain signals, and upper-cased letters with parenthesized ω, $A(\omega)$, denote the corresponding DTFT representations. Bold lower-cased letters, $\boldsymbol{a}$, represent column vectors. $*$ represents linear convolution. real$(\cdot)$ and imag$(\cdot)$ operators extract real and imaginary parts respectively. DFT$(\cdot)$ is the DFT operator. vec$(\cdot)$ stacks a discrete time/frequency domain signal into a column vector. diag$(\cdot)$ returns a diagonal matrix whose diagonal elements are the discrete time/frequency domain signal inside the bracket. Superscript ^{T} denotes matrix transpose, and superscript ^{H} denotes conjugate transpose. $\{\cdot\}_t$ with subscript t denotes a collection of variables indexed by t.

2. THE SOURCE-FILTER MODEL FOR PAT

2.1. The Signal Model

The signal model of PAT3 is similar to PAT2 [8]. It is based on the classical source-filter model, where the source is a mixture of glottal vibration and breathy noise, and the filter is the vocal tract

response. Formally, suppose each voiced speech frame, $s[t]$, is perfectly stationary. Modeling quasi-stationarity will be discussed in the next section. Then, $s[t]$ can be represented as

$$s[t] = (ae_v[t] + be_u[t]) * h[t] \tag{1}$$

where $e_v[t]$ and $e_u[t]$ are voiced and unvoiced excitation respectively. $h[t]$ is the impulse response of the vocal tract transfer function.

The unvoiced excitation is assume to be white Gaussian noise with unit variance.

$$\{e_u[t]\}_t \overset{iid}{\sim} \mathcal{N}(0,1) \tag{2}$$

The voiced excitation is a periodic signal, with each harmonic component modulated by $G(\omega)$, the transfer function of glottal source:

$$e_v[t] = \sum_d \text{real}\left[G(d\omega_0)e^{jd\omega_0(t-\tau)}\right] \tag{3}$$

where ω_0 is the fundamental frequency, also known as pitch frequency, and τ is the group delay.

We adopt the well-known three-pole model [11] for $G(\omega)$:

$$G(\omega) = \left[\left(1 - 2g_1\cos\beta e^{-j\omega} + g_1^2 e^{-2j\omega}\right)\left(1 - g_2 e^{-j\omega}\right)\right]^{-1} \tag{4}$$

where g_1, β and g_2 are the magnitude and phase of a maximum-phase pole pair, and the magnitude of a minimum-phase real pole.

The vocal tract system $H(\omega) \triangleq \text{DTFT}[h[t]]$ is modeled as a causal system [12], which can be well modeled by a few complex cepstral coefficients truncated at positive low quefrency [13]

$$H(\omega) = \exp\left(\text{DTFT}[\hat{h}[\hat{t}]]\right) \tag{5}$$

where $\hat{h}(\hat{t})$ is the truncated complex cepstrum.

2.2. The Probabilistic Model

PAT builds a probabilistic generative model based on the signal model described in eqs. (1) to (5). Formally, rewrite (1) as

$$s[t] = v[t] + u[t] \tag{6}$$

where

$$v[t] = ae_v[t] * h[t], \quad u[t] = be_u[t] * h[t] \tag{7}$$

are voiced and unvoiced portion of speech respectively. Define

$$\boldsymbol{v} = \text{vec}[\text{DFT}[v[t]]], \quad \boldsymbol{u} = \text{vec}[\text{DFT}[u[t]]] \tag{8}$$

and hidden variables

$$\boldsymbol{Z}_1 \triangleq \left\{a, b, \omega_0, \tau, g_1, g_2, \beta_2, \{\hat{h}(\hat{t})\}_{\hat{t}}\right\} \tag{9}$$

Then the probabilistic model essentially involves specifying the probability distribution of $\boldsymbol{v}$ and $\boldsymbol{u}$ conditional on $\boldsymbol{Z}_1$.

For $\boldsymbol{u}$, it can be derived from eqs. (2), (7) and (8) that real$[\boldsymbol{u}]$ are imag$[\boldsymbol{u}]$ are mutually independent and identically distributed as

$$\mathcal{N}(\boldsymbol{0}, b^2\text{diag}\left[|\text{DFT}[h[t]]|^2\right]) \tag{10}$$

except for the first and last (if DFT length is even) elements of imag$[\boldsymbol{u}]$, which are strictly 0.

For $\boldsymbol{v}$, the randomness is mainly from AM-FM effect. Its conditional distribution will be derived in the next section.

3. THE AM-FM MODEL FOR VOICED SPEECH

3.1. Adapted Baysian Spectral Estimation Model

If the voiced speech were stationary, the spectrum of voiced energy $v[t]$ is simply a weighted sum of sinusoids at multiples of pitch frequency. However, with AM/FM effect, the voiced speech can be instead represented as

$$v[t] = \sum_d \text{real}\left[\alpha_d\eta_d[t]\exp\left(jd\omega_0 t + jd\phi[t]\right)\right] \tag{11}$$

where

$$\alpha_d = aH(d\omega_0)G(d\omega_0)\exp(-jd\omega_0\tau) \tag{12}$$

is the complex transfer function at d-th harmonic. $\eta_d[t]$ denotes the real multiplicative variation (i.e. AM) of the transfer function at d-th harmonic. $\phi[t]$ denotes the random phase variation (i.e. FM) induced by pitch variation. We assume the phase variation at d-th harmonic is d times that at pitch frequency, thus giving $d\phi[t]$ in (11). Expanding (11), we get

$$v[t] = \sum_d \boldsymbol{x}_d[t]^T\boldsymbol{\xi}_d[t] \tag{13}$$

where

$$\boldsymbol{x}_d[t] = \begin{bmatrix} |\alpha_d|\cos(d\omega_0 t + \angle\alpha_d) \\ |\alpha_d|\sin(d\omega_0 t + \angle\alpha_d) \end{bmatrix} \tag{14}$$

which is essentially the vector of strictly periodic signal, and

$$\boldsymbol{\xi}_d[t] = \begin{bmatrix} \eta_d[t]\cos(d\phi[t]) \\ \eta_d[t]\sin(d\phi[t]) \end{bmatrix} \tag{15}$$

which is essentially the vector of AM-FM random variations. $|\cdot|$ and $\angle$ denote magnitude and angle of a complex number, respectively,

In Bayesian Spectral Estimation (BSE) [10], if $d\phi[t]$ is uniformly distributed, $\boldsymbol{\xi}_d[t]$ can be modeled as a multivariate Gaussian with zero mean and identity covariance matrix. However, uniform distribution of $d\phi[t]$ is not a reasonable assumption. Nevertheless, $\boldsymbol{\xi}_d[t]$ can still be reasonably approximated by a joint Gaussian with matched first and second moments, as will be shown in the next subsections.

3.2. The Autoregressive Model of $\boldsymbol{\xi}_d[t]$

Similar to BSE, the slowly time varying $\boldsymbol{\xi}_d[t]$'s are modeled as a first-order autoregressive process:

$$\boldsymbol{\xi}_d[t] = \lambda_d\boldsymbol{\xi}_d[t-1] + \boldsymbol{\varepsilon}_d[t] \tag{16}$$

With some assumptions[1], it can be shown that $\boldsymbol{\varepsilon}_d[t]$ can be reasonably assumed to satisfy independent Gaussian distribution:

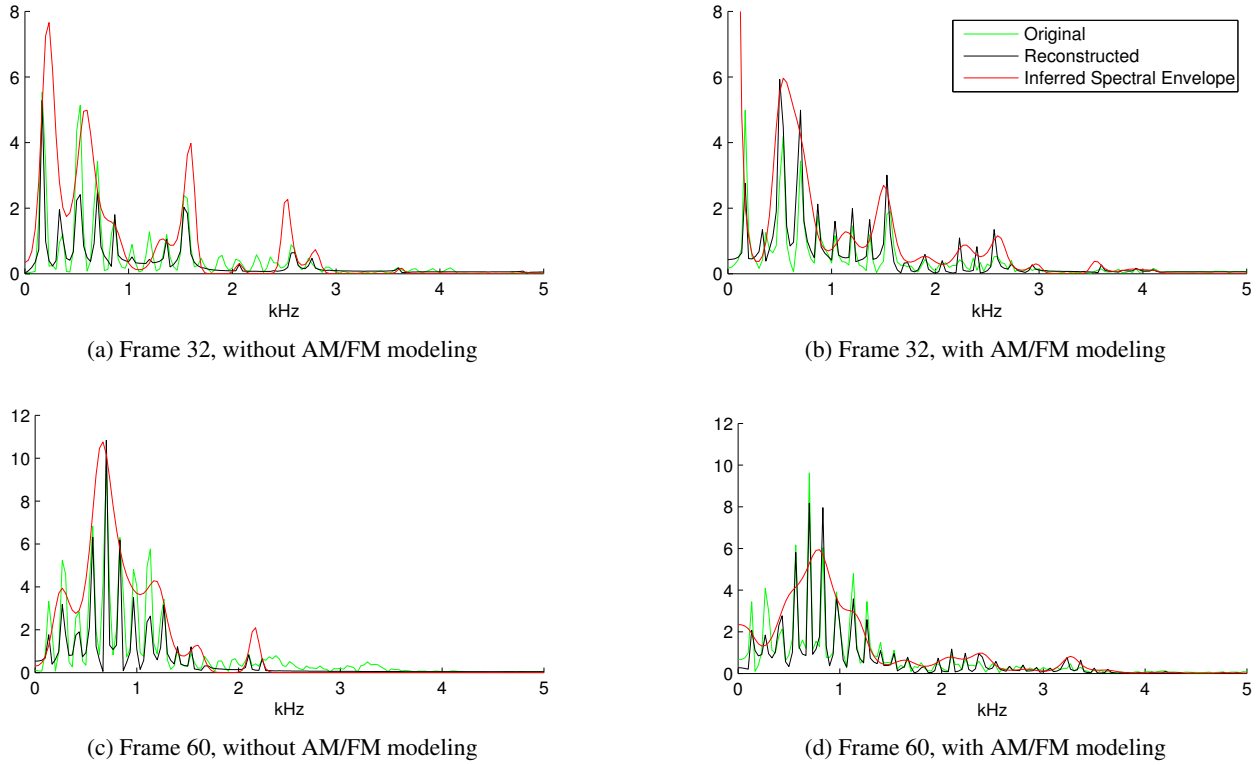
$$\boldsymbol{\varepsilon}_d[t] \sim \mathcal{N}\left(\boldsymbol{0}, \sigma_\varepsilon^2\begin{bmatrix} 1 & 0 \\ 0 & \rho_d^2 \end{bmatrix}\right) \tag{17}$$

where ρ_d is the ratio of standard deviations.

For the reason that would be clear below from (22), we assume that $\phi[t]$ satisfies independent increment process with Cauchy distributed increments. Then, it can be shown with some assumptions that λ_d decreases exponentially with d, i.e.

$$\lambda_d = \exp(-d\delta) \tag{18}$$

[1](i) The distribution of $\eta_d[t]$ is symmetric and centered at 0; (ii) $\phi[t]$ is small with respect to π, and has symmetric and unimodal distribution centered at 0.



(a) Frame 32, without AM/FM modeling

(b) Frame 32, with AM/FM modeling

(c) Frame 60, without AM/FM modeling

(d) Frame 60, with AM/FM modeling

Figure 1: The magnitude spectrum of the reconstructed voiced speech (black line) against the original magnitude spectrum (green line). PAT3 (right panels) is able to reclaim much of the voiced energy overlooked by PAT2 (left panels), especially in frequency band above 2kHz.

where δ is the parameter of the Cauchy-distributed increment of $\phi[t]$. Eq (18) is intuitively reasonable - as d goes up, the AM/FM variables $\boldsymbol{\xi}_d[t]$ become more random, and subsequently λ_d becomes closer to 0.

3.3. The Stationary Distribution of $\boldsymbol{\xi}_d[t]$

By quasi-stationarity of speech, it is reasonable to assume that the autoregressive process in (16) is close to stationary distribution, and it can be shown that the stationary distributions of $\boldsymbol{\xi}_d[t]$ is

$$\boldsymbol{\xi}_d[t] \sim \mathcal{N}\left(\mathbf{0}, \sigma_{\xi,d}^2 \begin{bmatrix} 1 & 0 \\ 0 & \rho_d^2 \end{bmatrix}\right) \tag{19}$$

with $\sigma_{\xi,d}$ determined by

$$\sigma_{\xi,d} = \frac{\sigma_\varepsilon}{\sqrt{1-\lambda_d^2}}, \tag{20}$$

Next, we derive an explicit relation of ρ_d depending on d. From (15), we have

$$\mathrm{pv}[d\phi[t]] = \arctan\left(\frac{\xi_d^{(2)}[t]}{\xi_d^{(1)}[t]}\right) \tag{21}$$

where $\mathrm{pv}[d\phi[t]]$ is the principal value of $d\phi[t]$. Thus, we can know the distribution of the warped phase $\mathrm{pv}[d\phi[t]]$ from the distribution of $\arctan\left[\xi_d^{(2)}[t]/\xi_d^{(1)}[t]\right]$. From this constraint, it can be further shown that a compatible distribution for the unwarped phase $d\phi[t]$ is a Cauchy distribution as follows

$$p_{d\phi[t]}(\varphi) = \frac{1}{\pi\gamma_d} \cdot \frac{\gamma_d^2}{\varphi^2 + \gamma_d^2} \tag{22}$$

where the parameter γ_d satisfies

$$\gamma_d = \frac{1}{2}\log\left(\sqrt{\frac{1+\rho_d}{1-\rho_d}}\right) \tag{23}$$

By the scaling property of Cauchy distribution[2], we have

$$\gamma_d = d\gamma_1 \tag{24}$$

Finally, combining (23) and (24) we obtain

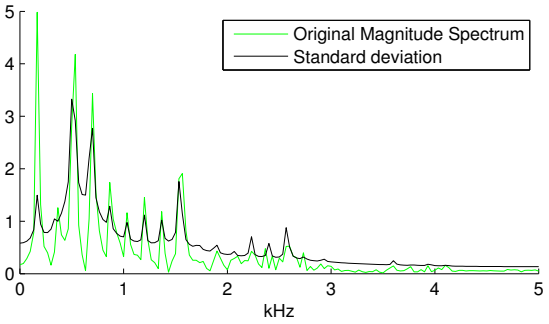
$$\rho_d = \tanh(2d\gamma_1) \tag{25}$$

which is intuitively reasonable - if d is small, $d\phi[t]$ is close to 0, then from (15), the variance of the second element of $\boldsymbol{\xi}_d$ is close to 0, and subsequently ρ_d is close to 0. On the other hand, if d is large, $\mathrm{pv}[d\phi[t]]$ will approach uniform distribution, and subsequently ρ_d will approach 1 (The model becomes standard BSE).

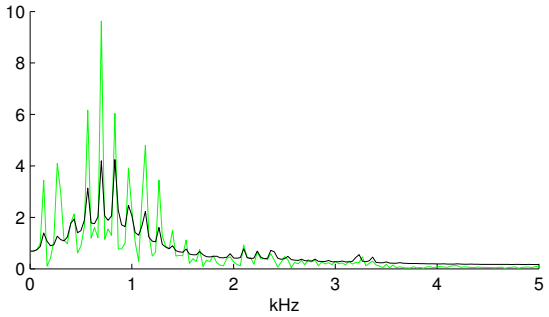
3.4. Model Summary

To sum up, the model of $\boldsymbol{v}$ is given by eqs. (13) to (18) and (25). The joint distribution of $\boldsymbol{v}$ is rather involved, but essentially it is a

[2]If $\phi[t] \sim \mathrm{Cauchy}(\gamma_1)$, then $d\phi[t] \sim \mathrm{Cauchy}(d\gamma_1)$.



(a) Frame 32



(b) Frame 60

Figure 2: Modeled standard deviation of each frequency bin of the magnitude spectrum (black line) against original magnitude spectrum, $\sqrt{\text{diag}(\boldsymbol{\Sigma})}$ (green line). AM/FM model admits widening of pitch pulse by adding variance around each pulse.

complex Gaussian with mean zero,

$$\boldsymbol{v} \sim \mathcal{CN}(\boldsymbol{0}, \boldsymbol{\Sigma}, \boldsymbol{C}) \tag{26}$$

where $\boldsymbol{\Sigma} = \text{E}(\boldsymbol{v}\boldsymbol{v}^H)$ and $\boldsymbol{C} = \text{E}(\boldsymbol{v}\boldsymbol{v}^T)$ are determined by speech signal $\{\boldsymbol{x}_d[t]\}$ modulated by the stochastic behavior of AM-FM variation $\{\boldsymbol{\xi}_d[t]\}$. Signal modeling with the covariance matrix of the (noisy) observation instead of modeling with the mean is a useful approach in source separation [14].

The hidden variables consist of two sets, the set $\boldsymbol{Z}_1$ that governs the signal $\{\boldsymbol{x}_d[t]\}$ as given by (9), and the set $\boldsymbol{Z}_2$ that governs the AM-FM variation $\{\boldsymbol{\xi}_d[t]\}$, given by

$$\boldsymbol{Z}_2 = \{\gamma_1, \delta\} \tag{27}$$

Notice that σ_ε is not distinguishable with a and therefore merged into a.

4. EXPERIMENTS

4.1. Configuration

Experiments are conducted to demonstrate the capability of the new AM/FM model in reconstructing speech with heavy AM/FM effect. The sampling rate is 10kHz. Speech is segmented into 30ms frames with 10ms frame shift. All the figures demonstrated are from speaker 1, utterance 1 in the Edinburgh speech corpus [15]. For comparison, both PAT3 with AM/FM modeling and the PAT2 model without AM/FM modeling are applied to infer the hidden variables for each voiced frame and reconstruct voiced spectrum based on the inferred values. [8] provides more details on the reconstruction approach. Due to limited space, the inference method is not elaborated, but basically it is a MAP optimization algorithm with loose priors by applying Monte Carlo sampling and quasi-Newton search. The dimension of $\hat{h}(\hat{t})$ is set to 26.

4.2. Reconstruction of Voiced Speech with Heavy AM/FM Effect

Voiced speech frames with significant AM/FM effect are studied. Figure 1 display the reconstructed magnitude spectrum of some voiced speech frames. The left panel is reconstructed by PAT2, the right by PAT3. The black line is reconstructed magnitude spectrum, green line the original magnitude spectrum, and red line the estimated spectral envelope, obtained by multiplying the inferred glottal source transfer function $G(\omega)$ and $H(\omega)$. An important observation is that in the original magnitude spectrum, the bandwidths of the harmonic pulses are small in low frequencies, and increase as frequency goes up. This widening of harmonic pulses is the major effect of AM/FM, and becomes more significant in frequency band above 2kHz, which agrees with (25) and (18).

As for the reconstruction accuracy, PAT2 significantly underestimates voiced energy in frequency band above 2kHz. This is because PAT2 does not account for the widening of the harmonic pulses, and ascribes this variation to unvoiced energy. On the other hand, PAT3 is able to more accurately estimate the spectral envelope.

4.3. Standard Deviation of v

As mentioned in section 3.4, the new PAT models the speech signal as a zero mean Gaussian. Information of the modeled signal is incorporated in the covariance matrix. To give a better idea of how it models AM/FM effect, fig. 2 shows the squared root of the diagonal the covariance matrix , $\sqrt{\text{diag}(\boldsymbol{\Sigma})}$, of the magnitude spectrum (black line) as in (26), against the original magnitude spectrum (green line). As can be seen, PAT3 allows some variations around each harmonic pulse, and the variation are larger and of greater range as frequency goes up. This represents the essential mechanism of incorporating AM/FM effect, or say the widening of pitch pulses. According to (16), λ_d governs the range of variation. Since λ_d is determined by the hidden variable δ, which is inferred for each speech frame, the range of variation generally matches the width of pulse.

5. CONCLUSION AND FUTURE DIRECTION

This paper represents our ongoing progress to develop a complete probabilistic generative model of speech - PAT. In particular, we significantly improve voiced modeling of PAT by modeling AM/FM variations through adapting Bayesian Spectrum Estimation method. Preliminary studies have shown that it can infer hidden variables for voiced speech affected by AM/FM variations. One challenge of PAT, which impedes us from conducting large scale experiments, is the high computational complexity of inference. Currently the inference is conducted separately for each frame. Introducing smoothing transitions across consecutive frames to consider speech dynamics may reduce the search space to speed up, but with the demand for a high-performance sequential inference algorithm. We will work on this issue as one future direction.

6. REFERENCES

[1] H. Kameoka, T. Nishimoto, and S. Sagayama, "A multipitch analyzer based on harmonic temporal structured clustering," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 15, no. 3, pp. 982–994, 2007.

[2] J. P. Cabral, "Hmm-based speech synthesis using an acoustic glottal source model," Ph.D. dissertation, School of Informatics, The University of Edinburgh, 2011.

[3] U. Simsekli, J. Le Roux, and J. R. Hershey, "Non-negative source-filter dynamical system for speech enhancement," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2014, pp. 6206–6210.

[4] J. R. Hershey, S. J. Rennie, P. A. Olsen, and T. T. Kristjansson, "Super-human multi-talker speech recognition: A graphical modeling approach," *Computer Speech & Language*, vol. 24, no. 1, pp. 45–66, 2010.

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

[6] G. Degottex, P. Lanchantin, A. Roebel, and X. Rodet, "Mixed source model and its adapted vocal tract filter estimate for voice transformation and synthesis," *Speech Communication*, vol. 55, no. 2, pp. 278–294, 2013.

[7] Z. Ou and Y. 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.

[8] Y. Zhang, Z. Ou, and M. Hasegawa-Johnson, "Improvement of probabilistic acoustic tube model for speech decomposition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2014, pp. 7929–7933.

[9] T. Quariteri, *Discrete-time Speech Signal Processing*. Prentice-Hall, 2002, p. 64.

[10] Y. Qi, T. P. Minka, and R. W. Picara, "Bayesian spectrum estimation of unevenly sampled nonstationary data," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. IEEE, 2002, pp. II–1473.

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

[12] O. Fujimura, "Analysis of nasal consonants," *The Journal of the Acoustical Society of America*, vol. 34, p. 1865, 1962.

[13] A. V. Oppenheim, R. W. Schafer, J. R. Buck, *et al.*, *Discrete-time signal processing*. Prentice Hall Upper Saddle River, 1999, vol. 5.

[14] A. Ozerov, E. Vincent, and F. Bimbot, "A general flexible framework for the handling of prior information in audio source separation," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 20, no. 4, pp. 1118–1133, 2012.

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