IMPROVEMENT OF PROBABILISTIC ACOUSTIC TUBE MODEL FOR SPEECH DECOMPOSITION

Yang Zhang, Zhijian Ou, Mark Hasegawa-Johnson

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

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. Under this motivation, we have proposed 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

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 voiced excitation can be modeled as periodic repetition of the unvoiced excitation, which is closer to the nature of speech. Therefore, PAT2 makes U/V states a continuum by introducing voiced 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 \( \omega \), e.g. \( X(\omega) \), to denote its DTFT; bold lower-case letter, e.g. \( \mathbf{x} \), for column vectors and bold upper-case letter, e.g. \( \mathbf{X} \), for matrices; IDTFT\[^{-1}\] for inverse DTFT operator; \( \ast \) for circular convolution.

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 voiced excitation can be modeled as periodic repetition of the unvoiced excitation, which is closer to the nature of speech. Therefore, PAT2 makes U/V states a continuum by introducing voiced 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 \( \omega \), e.g. \( X(\omega) \), to denote its DTFT; bold lower-case letter, e.g. \( \mathbf{x} \), for column vectors and bold upper-case letter, e.g. \( \mathbf{X} \), for matrices; IDTFT\[^{-1}\] for inverse DTFT operator; \( \ast \) for circular convolution.

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 \( \omega \), e.g. \( X(\omega) \), to denote its DTFT; bold lower-case letter, e.g. \( \mathbf{x} \), for column vectors and bold upper-case letter, e.g. \( \mathbf{X} \), for matrices; IDTFT\[^{-1}\] for inverse DTFT operator; \( \ast \) for circular convolution.

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 \( \omega \), e.g. \( X(\omega) \), to denote its DTFT; bold lower-case letter, e.g. \( \mathbf{x} \), for column vectors and bold upper-case letter, e.g. \( \mathbf{X} \), for matrices; IDTFT\[^{-1}\] for inverse DTFT operator; \( \ast \) for circular convolution.

frequency domain, it is represented as
\[ V_t(\omega) = G_t(\omega) e^{-j\omega \tau_\omega} \sum_k \delta(\omega - k\omega_{0\omega}) \]  
where \( G_t(\omega) \) is the DTFT of glottal wave of one cycle; \( \delta(\omega) \) is the dirac delta function; \( \omega_{0\omega} \) is the fundamental frequency in radians; \( \tau_\omega \) 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:
\[ u_t[n] = \text{IDTFT}[U_t(\omega)] \buildrel \text{iid} \over \sim N(0, 1) \]  
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 effects 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
\[ G_t(\omega) = \frac{1}{(1 + 2g_{1t} \cos \beta_i e^{-j\omega} + g_{2t}^2 e^{-2j\omega})(1 + g_{2t} e^{-j\omega})} \]  
where \( g_{1t}, \beta_i \) 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
\[ \hat{h}_t[n] = \text{IDTFT}[\log(H_t(\tilde{\omega}))] \]  
where \( \tilde{\omega} \) is mel-frequency:
\[ \tilde{\omega} = m(\omega) = \begin{cases} \omega/2000 + 1 \times 2254\pi & \omega < 2000\pi \\ \text{otherwise} & \end{cases} \]  
According to previous study \[9\], vocal tract can be well modeled by a minimum phase system. Thus, it can be proved \[10\] that \( \hat{h}_t[n] \) is right-sided, namely 0 when \( n < 0 \), and if group delay and the sign of gain of \( H(\omega) \) are properly removed, \( \hat{h}_t[n] \) decays at the rate of \( 1/n \). According to (1) and (2), the sign of gain is controlled by \( a_t \) and group delay by \( \tau_\omega \). Therefore, we can use \( \hat{h}_t[n], 0 < n \leq K \) for small \( K \), named MFC\(^3\), to represent \( H_t(\omega) \):
\[ H_t(\omega) = \exp \left( \sum_{n=1}^{K} \hat{h}_t[n] \exp(-jm(\omega)\hat{n}) \right) \]  
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 \( \Theta \) is
\[ \Theta = \bigcup_t \left\{ \{a_t, b_t, \omega_{0t}, \tau_\omega, g_{1t}, \beta_i, g_{2t}\} \cup \{\hat{h}_t[n]\} \right\} \]  

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
\[ s_t = \begin{bmatrix} S_t^{(r)}(2\pi 0 N) \ldots S_t^{(r)}(2\pi (N/2) N) \\ S_t^{(i)}(2\pi (N/2 - 1) N) \ldots S_t^{(i)}(2\pi (N/2) N) \end{bmatrix}^T \]  
where superscript \((r)\) and \((i)\) denotes real part and imaginary part respectively. \( S_t^{(1)}(0) \) and \( S_t^{(2)}(\pi) \) 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 \( s_t \) to minimize the error. According to (1), we have
\[ \tilde{s}_t \equiv F s_t = a_t F \xi_t + b_t F \eta_t \]  
where \( F \) is a matrix containing rows of overlapping triangular windows, whose end points are uniform in mel-scale; \( \xi_t \) and \( \eta_t \) are compact real DFT representation of \( V_t(\omega) H_t(\omega) \oplus W_t(\omega) \) and \( U_t(\omega) H_t(\omega) \oplus W_t(\omega) \) respectively.

\( \eta_t \) 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, \( \eta_t \) can be further represented as
\[ \eta_t = H_t u_t \]  
where \( u_t \) is the compact real DFT representation of \( U_t(\omega) \) and \( H_t \) 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 \( u_t \) is a standard multivariate Gaussian variable, i.e. with zero mean and identity covariance matrix.

Therefore
\[ \tilde{s}_t \sim N \left( a_t F \xi_t, b_t^2 FH_tH_t^T F^T \right) \]  
which defines the likelihood of the observed speech frame.
In this section, we display a set of preliminary results 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].

In this experiment, we set $\sigma_{\omega_0}^2 = 1, \sigma_a^2 = \sigma_b^2 = 0.01, \sigma^2_{\hat{v}} = \sigma^2_{\hat{w}} = 0.1, \sigma_{\hat{v}}^2 = 1e-5$.

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

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].

In this experiment, we set $\sigma_{\omega_0}^2 = 1, \sigma_a^2 = \sigma_b^2 = 0.01, \sigma^2_{\hat{v}} = \sigma^2_{\hat{w}} = 0.1, \sigma_{\hat{v}}^2 = 1e-5$.

In this experiment, we set $B = 50, C = 10$.
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 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, $\tau_i$ 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 $\tau_i + 2k\pi/\omega_0$, 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.

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 S-TRAIGHT 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


