Probabilistic acoustic tube: a probabilistic generative model of speech for speech analysis/synthesis

Zhijian Ou
ozj@tsinghua.edu.cn

Yang Zhang
zhangyangbill@gmail.com

Department of Electronic Engineering, Tsinghua University, Beijing 100084, China

Abstract

Most speech analysis/synthesis systems are based on the basic physical model of speech production - the acoustic tube model. There are two main drawbacks with current speech analysis methods. First, a common design paradigm seems to build a special-purpose signal-processing front-end followed by (when appropriate) a back-end based on probabilistic models. A difficulty is that most features are nonlinear operators of the speech waveform, whose statistical behavior is hard to be modeled. Second, different tasks of speech analysis are carried out separately. These practices are admittedly useful but not optimal due to the incomplete use of available information. These examinations motivate us to directly model the spectrogram and to integrate together the three fundamental speech parameters - the pitch, energy and spectral envelope. We successfully devise such a model called probabilistic acoustic tube (PAT) model. The integration is performed in a principled manner with explicit physical meaning. We demonstrate the capability of PAT for a number of speech analysis/synthesis tasks, such as pitch tracking under both clean and additive noise conditions, speech synthesis, and phoneme clustering.

1 Introduction

Speech analysis/synthesis refers to a family of speech processing applications, such as speech modification, coding, enhancement, and recognition (Quatieri, 2001). Most speech analysis/synthesis systems are based on the basic physical model of speech production - the acoustic tube model, also known as the source-filter model (Quatieri, 2001). Speech is viewed as the result of passing the glottal excitation source through the vocal tract, which in a short-time interval could be represented as an acoustic tube with a fixed shape and further be modeled as a linear time-invariant system, or say, a filter. The excitation function could be either a quasi-periodic pulse train (for voiced speech) or a random noise (for unvoiced sounds). A particular sound is completely characterized by the excitation function, the excitation gain and the filter exercised in the speech production. These correspond to three important parametric representation of speech – the pitch, energy and spectral envelope respectively.

In analysis, we take apart the speech waveform to extract underlying parameters and possible further high-level information, e.g. the phoneme being uttered and the speaker identity. In synthesis, after some desirable transformation, the estimated parameters are put together to reconstruct the waveform. There are two main drawbacks with current speech analysis methods.

First, a common design paradigm seems to build a special-purpose signal-processing front-end followed by (when appropriate) a back-end based on probabilistic models. The purpose of the front-end is to extract the most relevant features for the target task. Two widely-used signal-processing techniques to extract spectral envelopes are LPC (linear predictive coding) and cepstrum. Most pitch estimation algorithms first extract a set of nonlinear front-end features (e.g. the normalized autocorrelation) that exhibit special behavior when voice speech is uttered and then model this behavior to track pitch. High-level analysis such as speech recognition uses hidden Markov models (HMMs) to model only the spectral envelopes, which are known to be directly related to the speech sounds being uttered and are parameterized by cepstrum features in the front-end. In this paradigm, We need to deal with two difficult problems of finding the most relevant features and building more powerful probabilistic models to accommodate the randomness of the features. A difficulty is that most features are nonlinear operators of the speech waveform, whose statistical behavior is hard to be modeled.

In this paper, we investigate to directly model the spectrogram that is a fundamental and linear representation of speech. This has two potential advantages. First, in most real-world applications (e.g. in the cocktail party scenario), this will preserve additivity and make it possible to perform robust analysis in the presence of multiple sound sources that mix additively. For example, in Bach and Jordan (2005), the models of spectrogram for each speaker are joined together through factorial HMM modeling to

Appearing in Proceedings of the 15th International Conference on Artificial Intelligence and Statistics (AISTATS) 2012, La Palma, Canary Islands. Volume XX of JMLR: W&CP XX.

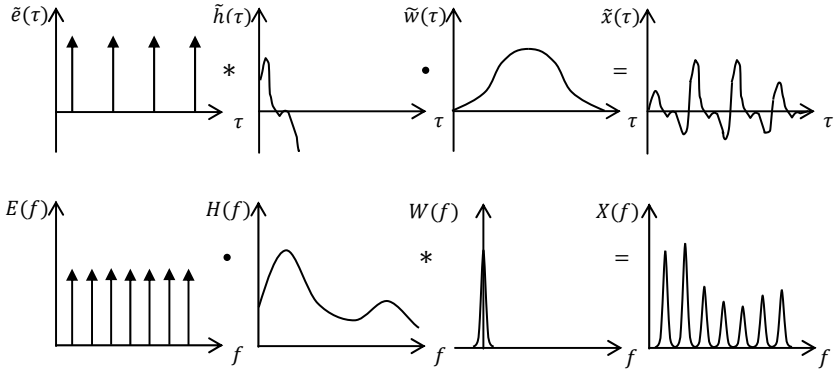


Figure 1: Signal-processing modeling of a speech frame. The upper graphs are in time domain, and the lower graphs are in magnitude spectral domain. The dot and the star represent the multiplication and convolution operation respectively.

achieve multiple pitch tracking. Second, note that according to information theory, cascaded processing will reduce the information. So if we improve modeling of the original spectrogram, we will obtain better performances for various speech analysis tasks.

Next we examine the second drawback with current speech analysis methods. Note that different tasks of speech analysis are carried out separately. The pitch and spectral envelope are usually estimated separately. Speech recognition is performed by HMM-based modeling of only the spectral envelopes. In the most recent model-based approaches to compensate for additive and convolutional noises in speech recognition, e.g. using vector Taylor series expansion (Li et al., 2009), the pitch information is still ignored which is known to be an important clue to discriminate speech from noise. The above practices are admittedly useful and effort-saving, but not optimal due to the incomplete use of available information. It has been noted in Kameoka et al. (2010) that estimation of the pitch and of the envelope has a chicken-and-egg relationship and should be performed jointly. The more reliable the pitch determination is the more accurate the envelope estimation becomes, and vice versa. In Stephenson et al. (2004), it is observed that the cepstral-based features are sensitive to "auxiliary" information, such as pitch, energy, etc. The practice of incomplete speech analysis may be partly because it remains a long-standing problem to construct a unified probabilistic model to integrate the three fundamental speech parameters - the pitch, energy and spectral envelope, beyond the physical acoustic tube model.

The above discussions motivate us to propose a probabilistic generative model of the spectrogram based on the physical acoustic tube modeling of speech production. We successfully devise such a model called probabilistic acoustic tube (PAT) model in the sense that the excitation function, the excitation gain and the transfer function that models the resonant characteristics of the acoustic tube are all probabilistically modeled. The integration is performed in a principled manner with explicit physical meaning. When we describe the PAT in the graphical modeling framework, it clearly shows how the pitch, energy and spectral envelope are interacted to generate the spectrogram.

We could learn a PAT from speech utterances in a supervised or unsupervised way using the EM algorithm. Once we have such a probabilistic generative model of speech, the observed spectrogram can be interpreted/analyzed by performing inference over hidden variables, such as the pitch, the uttered phoneme, etc. On the other hand, the inferred values or trained parameters can be used to reconstruct the speech. Our study in this paper is in spirit similar to the generative modeling approach to computer vision (Frey 1999) that successfully accounts for different sources of variability in images and relies on learning and inference to perform various image analysis tasks. We demonstrate the capability of PAT for a number of speech analysis/synthesis tasks, such as pitch tracking under both clean and additive noise conditions, speech synthesis, and phoneme clustering.

Notations. We use the lower case symbols with hats, e.g. $\tilde{x}(\tau)$, to denote the time-domain signals over continuous time τ. The corresponding capital symbols, e.g. $X(f)$ and $X(\omega)$, represent the Fourier transforms over continuous-time frequency f and discrete-time frequency ω respectively. For speech frame t, the discrete Fourier transform defined by sampling $X(\omega)$ uniformly over N discrete-frequency bins is denoted as x_t.

2 Probabilistic Acoustic Tube (PAT)

2.1 Signal-processing modeling

According to the physical acoustic tube modeling of speech production, a frame of the speech signal $\tilde{x}(\tau)$ can be modeled as the windowed convolution of the vocal tract impulse response $\tilde{h}(\tau)$ with the source excitation $\tilde{e}(\tau)$, as shown in Figure 1:

$$\tilde{x}(\tau) = \left[\tilde{e}(\tau) * \tilde{h}(\tau)\right] \cdot \tilde{w}(\tau) \quad (1)$$

Here τ is time and $\tilde{w}(\tau)$ is the window function. The dot and the star represent the multiplication and convolution operation respectively. In the discrete-time Fourier domain,

we have

$$X(\omega) = [E(\omega)H(\omega)] * W(\omega) \quad (2)$$

where ω is the discrete-time frequency. $E(\omega)$, $H(\omega)$ and $W(\omega)$ are the Fourier transforms for the excitation function, the vocal tract filter and the window function respectively.

For voiced speech, the excitation function $\tilde{e}(\tau)$ could be modeled as a pulse sequence, and its Fourier transform is again a pulse sequence:

$$E_l(\omega) = \sum_n \delta(\omega - n\eta_l) \quad (3)$$

Here $\delta(\cdot)$ is the Dirac function and n runs over the integers. The pitch is discretized to be equally spaced in the midi number scale with a total of L elements. The mapping from Hz to midi number is $\text{midi} = 69 + 12\log_2(\text{Hz}/440)$. $E_l(\omega)$ is the excitation spectrum for the l-th discretized pitch, denoted as η_l, $l = 1, \dots, L$. Then, the magnitude spectrum of the voiced speech with the pitch frequency η_l can be approximated as an amplitude-modulated comb:

$$\begin{aligned} |X(\omega)| &= |[E_l(\omega)H(\omega)] * W(\omega)| \\ &\approx \sum_n |H(n\eta_l)||W(\omega - n\eta_l)| \end{aligned} \quad (4)$$

The approximation is justified under the condition that the magnitude spectrum of the sum of multiple signal components is approximately equal to the sum of the magnitude spectra of these components. The smaller the spectral leakage from adjacent components, which means the cross term $|H(m\eta_l)W(\omega - m\eta_l)||H(n\eta_l)W(\omega - n\eta_l)|$ with $m \neq n$ is sufficiently small, the higher the accuracy of this approximation. Considering that the spectrum of the window function usually has a narrow main lobe and low side lobes, the above approximation error is low.

For unvoiced speech, the white noise is usually used as the excitation function $\tilde{e}(\tau)$, which has been shown to produce satisfactory performance for speech synthesis. Considering that the expected magnitude spectrum of the white noise is constant, we approximate the excitation spectrum $E(\omega)$ for unvoiced speech by a constant spectrum. Then, the magnitude spectrum of the unvoiced speech can be approximated as:

$$\begin{aligned} |X(\omega)| &= \left| \sum_\xi H(\xi)W(\omega - \xi) \right| \\ &\approx \sum_\xi |H(\xi)||W(\omega - \xi)| \end{aligned} \quad (5)$$

In conclusion, based on the acoustic tube modeling of speech production, we obtain Equ. (4) and (5) that are the signal-processing models of the magnitude spectra for voiced and unvoiced speech respectively. The approximation errors involved above will be accounted for in the following probabilistic modeling.

2.2 Probabilistic modeling

Based on the above signal-processing models of speech, we define the following generative graphical model for an utterance, as shown in Figure 2. We introduce five random variables for each speech frame t.

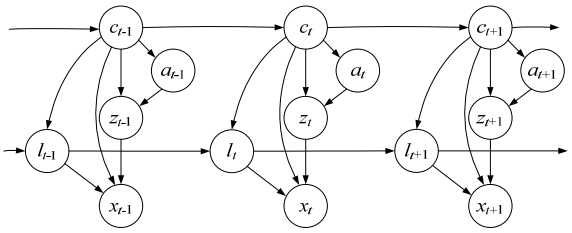


Figure 2: Probabilistic graphical model of PAT

First, each speech frame is supposed to be in one of a finite number of phonetic states as indicated by c_t. The phonetic state variable c_t can take on a total number of $K_v + K_u$ states. The number of voiced and unvoiced/silence phonetic states are K_v and K_u respectively. In supervised learning, the phonetic states can be assigned to represent phonemes. In unsupervised learning, the speech frames are clustered into $K_v + K_u$ phonetic clusters. In this paper, we use the following simple conditional distribution $p(c_t|c_{t-1})$ to constrain the unvoiced-voiced transition:

$$p(c_t|c_{t-1}) = \begin{cases} Z\exp(-\alpha) & \text{if there is a u-v transition} \\ Z & \text{otherwise} \end{cases} \quad (6)$$

where α is an empirically determined parameter and Z is the normalization constant.

Second, l_t represents the discretized pitch frequency at frame t. $l_t = 0$ indicates that unvoiced excitation is used, while $l_t = 1, \dots, L$ indicates that voiced excitation is used. In order to prevent abrupt changes in pitch, we constrain $p(l_t|l_{t-1}, c_t)$ to be zero if the pitch frequency difference between adjacent frames exceeds 5 midi numbers, or if l_t and c_t fall in different unvoiced/voiced categories.

Third, there is a scalar variable a_t to represent the excitation gain, which is assumed to be Gaussian distributed as follows, depending on c_t:

$$p(a_t|c_t = c) = \mathcal{N}(a_t; m_c, \sigma_c^2) \quad (7)$$

Explicit modeling of the excitation gain is beneficial for exploiting energy information. For example, unvoiced frames usually have lower energy than voiced frames.

Fourth, a M-dimenional variable z_t is used to represent the first M discrete cosine transform (DCT) coefficients of the vocal tract spectral envelope h_t, which is defined by sampling $H(\omega)$ uniformly over N discrete-frequency bins . We have

$$h_t = Cz_t \quad (8)$$

where C is the inverse DCT matrix of the size $N \times M$. The DCT is used here to ensure that the vocal tract spectral envelope is smoothly modeled. The vocal tract spectral envelope is governed by the phonetic state variable c_t and

is assumed to be Gaussian distributed as follows, depending on c_t:

$$p(z_t|c_t = c, a_t) = \mathcal{N}(z_t; a_t\mu_c, \Phi_c) \quad (9)$$

Here a_t is the excitation gain defined in Equ. (7), and μ_c is M-dimensional column vector subject to the following normalizing constraint, where we use the apostrophe to represent transpose:

$$\mu_c'\mu_c = 1 \quad (10)$$

The reason for the normalization is that it allows the amplitude of speech to be explicitly modeled. In this way, speech frames with closely shaped spectral envelopes and different energies could be modeled as different realizations of the same phoneme. Otherwise the amplitude information would be absorbed into the spectral envelopes. Φ_c is the full covariance matrix to enforce strong constraints on the spectral envelope.

Fifth, the observed magnitude spectrum x_t is defined by sampling $X(\omega)$ uniformly over N discrete-frequency bins. The N-dimensional vector x_t is modeled as follows:

$$x_t = E_{l_t}h_t + n_t \quad (11)$$

By regarding Equ. (4) and (5) as the result of multiplying a matrix with a vector, we have the following definition of the excitation matrix E_{l_t} (abbreviated as E_l below) of the size $N \times N$.

Recall that the l-th discretized pitch frequency is denoted as η_l. $l = 1, \ldots, L$ indicates that voiced excitation is used. In this case, E_l has non-zero columns only at the harmonics of the pitch frequency η_l. Specifically, for the m-th harmonic of the pitch frequency η_l, we use $j = \text{argmin}_k|v(k) - m\eta_l|$ to determine the specific j-th column of E_l, which is a bump centered at frequency $m\eta_l$, defined as the Fourier transform of the window function. m runs over the integers from 1 to the number of harmonics for pitch η_l in the bandwidth concerned. $v(k)$ represents the frequency of the k-th bin used in discrete Fourier transforms, $k = 0, \cdots, N-1$. To put the above together, we have the following definition for the element at the *i*-th row and *j*-th column of E_l:

$$(E_l)_{ij} = \begin{cases} W(v(i) - m\eta_l), & \text{if } j = \underset{k}{\text{argmin}}|v(k) - m\eta_l| \\ 0, & otherwise \end{cases} \quad (12)$$

An example of the voiced excitation matrices is shown in Figure 3(a). For $l_t = 0$ which indicates that unvoiced excitation is used, we have

$$(E_l)_{ij} = W\big(v(i) - v(j)\big) \quad (13)$$

An example of the unvoiced excitation matrix is shown in Figure 3(b).

In Equ. (11), n_t is the noise accounting for the approximating errors from Equ. (4) and (5), which is assumed to be Gaussian distributed as follows, depending on c_t:

$$p(n_t|c_t = c) = \mathcal{N}(n_t; 0, m_c^2\Psi) \quad (14)$$

where Ψ is a diagonal covariance matrix.

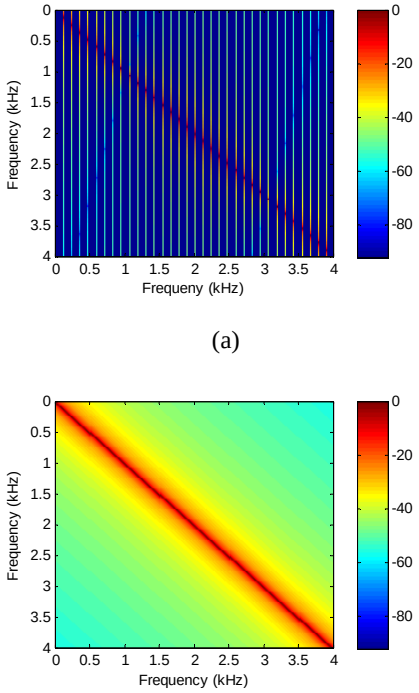


Figure 3: Examples of excitation matrix E_l (in dB).
(a) The particular E_l matrix for pitch frequency 259.3685Hz, which consists of shifted bumps in separated columns. Dark blue represents negative infinity in dB, namely zero in the original domain.
(b) The E_l matrix for unvoiced speech, which consists of shifted bumps in every column.

Finally, after defining $\Gamma_{l_t} = E_{l_t}C$, we have

$$x_t = \Gamma_{l_t}z_t + n_t \quad (15)$$

The joint probability distribution of a T-frame utterance is given as follows, where we use the Matlab notation to represent a set of variables, e.g. $x_{1:T} \triangleq x_1, \cdots, x_T$:

$$\begin{aligned} &p(c_{1:T}, l_{1:T}, z_{1:T}, a_{1:T}, x_{1:T}) \\ &= \prod_t p(c_t|c_{t-1})p(l_t|l_{t-1}, c_t) \\ &\cdot p(a_t|c_t)p(z_t|c_t, a_t)p(x_t|l_t, c_t, z_t) \\ &= \prod_t p(c_t|c_{t-1})p(l_t|l_{t-1}, c_t) \\ &\cdot \mathcal{N}(a_t; m_{c_t}, \sigma_{c_t}^2)\mathcal{N}(z_t; a_t\mu_{c_t}, m_{c_t}^2\Phi_{c_t})\mathcal{N}(x_t; \Gamma_{l_t}z_t, m_{c_t}^2\Psi) \end{aligned} \quad (16)$$

In conclusion, speech is probabilistically generated as

shown in Figure 2. To generate a speech frame, a phonetic state c_t is randomly chosen. The selected mean normalized vocal tract shape is multiplied with a random excitation gain a_t and then a random noise is added to create the actual vocal tract shape. Next, the excitation specified by a random pitch l_t is passed through the above randomly created vocal tract, and a certain amount of noise is added to produce the final observed frame.

2.3 Related works

It is worthwhile to remark on the novelty of PAT and its relationship between previous works.

At a first look at the state-of-the-arts of speech processing (e.g. pitch estimation, speech recognition, source separation and so on), it seems that there are generative models of speech. But most of them are actually generative models of the speech features (e.g. the correlogram or the cepstrum). We are interested in direct modeling of the spectrogram, whose advantages are discussed in the Introduction section. Therefore, in the following, we only compare PAT with other related works that directly model the spectrogram (Reyes-Gomez et al. 2005, Bach and Jordan 2005, Kameoka et al. 2006, Hershey et al. 2010).

The first distinctive feature of PAT, which is missing in these related works, is that PAT explicitly considers the energy information parameterized by the excitation gain a_t. The energy contour of an utterance contains important information about the phonetic identity of the sounds within the utterance. Moreover, when dealing with speech mixtures, the gain-level of each source component is an important cue for multi-pitch tracking and source separation.

The second distinctive feature of PAT, which is missing in these related works, is that the vocal tract response and the pitch are decoupled yet jointly modeled, and the probabilistic modeling of the spectral envelope is further augmented by introducing the underlying phoneme being uttered. In most previous works, the magnitude spectrum is simply directly modeled, for example, as a mixture of Gaussians (Hershey et al. 2010), where the effects of the vocal tract response and the pitch are mixed.

To put the two novel elements of PAT together, the key improvement over these state-of-the-arts is that the state space of PAT is meaningfully factored over phonetic state c_t, pitch l_t, and excitation gain a_t. The state space of these previous works is not as factored as in PAT, so that a large number of states are required. Although this could be remedied by adding some 'stickiness' bias, the model will end up with a 'blurry' set of states. This is also noted in Reyes-Gomez, et al., (2005).

2.4 Parameter estimation

In this section, we describe the EM algorithm to estimate the parameters of a PAT:

$$\Theta \triangleq \{\mu_c, \Phi_c, \Psi, m_c, \sigma_c^2\} \tag{17}$$

In the E-step, we perform the forward-backward algorithm to calculate the posterior of c_t and l_t:

$$p(c_t, l_t|x_{1:T}) \propto \alpha(c_t, l_t)\beta(c_t, l_t) \tag{18}$$

where

$$\alpha(c_t, l_t) \triangleq p(c_t, l_t, x_{1:t}) \tag{19}$$

$$\beta(c_t, l_t) \triangleq p(x_{t+1:T}|c_t, l_t) \tag{20}$$

To reduce computation cost, we apply Viterbi approximation. Moreover, it can be derived that both $\mathrm{p}(a_t|x_t, c_t, l_t)$ and $p(z_t|x_t, a_t, c_t, l_t)$ are Gaussian distributed. See the appendix in the supplementary material for details.

In the M-step, the parameters are re-estimated as follows:

$$\hat{\mu}_c = \left(\sum_t \gamma_{t,c}\langle a_t^2\rangle_{p(a_t|\, x_{1:T}, c_t=c)}\, \mathbb{I} + \, 2\lambda_c \hat{\Phi}_{\mathrm{c}} \mathbf{1}'\mathbf{1}\right)^{-1} \cdot \sum_t \gamma_{t,c}\langle a_t z_t\rangle_{p(a_t, z_t|x_{1:T}, c_t=c)} \tag{21}$$

$$\hat{\Phi}_{\mathrm{c}} = \frac{\sum_t \gamma_{t,c}\langle (z_t - a_t\hat{\mu}_c)(z_t - a_t\hat{\mu}_c)'\rangle_{p(z_t, a_t|x_{1:T}, c_t=c)}}{\sum_t \gamma_{t,c}} \tag{22}$$

$$\hat{\Psi} = \frac{1}{T}\sum_t \mathrm{diag}\left[\langle \hat{m}_{c_t}^{-2}\left(x_t - \Gamma_{l_t} z_t\right)\left(x_t - \Gamma_{l_t} z_t\right)'\rangle_{p(z_t, c_t, l_t|x_{1:T})}\right] \tag{23}$$

Here $\gamma_{t,c} \triangleq p(c_t = c|x_{1:T})$, $\mathbf{1} \triangleq \{1\}_{M\times 1}$, $\mathbb{I}$ denotes the identity matrix, and λ_c is the Lagrange multiplier. T denotes the total number of frames. The angle bracket $\langle\cdot\rangle_p$ represents the expectation operator with respect to the distribution p. $\mathrm{diag}[\cdot]$ represents extracting the diagonal elements of a matrix to form a diagonal matrix.

m_c is re-estimated by solving the following quartic equation, where $tr[\cdot]$ represents the trace of a matrix,

$$0 = \sum_t \langle \hat{\sigma}_c^{-2}(a_t - \hat{m}_c)\rangle_{p(c_t=c, a_t|x_{1:T})} - \left(M\sum_t \gamma_{t,c}\right)\hat{m}_c^{-1} + \sum_t \langle \hat{m}_c^{-3} tr\left[\left(x_t - \Gamma_{l_t} z_t\right)\left(x_t - \Gamma_{l_t} z_t\right)'\hat{\Psi}^{-1}\right]\rangle_{p(z_t, c_t=c, l_t|x_{1:T})} \tag{24}$$

$$\hat{\sigma}_c^2 = \frac{\sum_t \gamma_{t,c}\langle a_t^2 - 2a_t\hat{m}_c\rangle_{p(a_t|\, x_{1:T}, c_t=c)}}{\sum_t \gamma_{t,c}} + \hat{m}_c^2 \tag{25}$$

3 Experiments

In the experiments, we demonstrate that the PAT model can be used for a number of speech analysis/synthesis tasks, such as pitch tracking under both clean and additive noise conditions, speech synthesis, and phoneme clustering.

3.1 Pitch tracking

We evaluate pitch tracking performance on Edinburgh database (Bagshaw 1993), which consists of a male speaker and a female speaker, each producing 50 sentences. The ground truth pitch labels are provided based on a simultaneously recorded signal of the

laryngograph. We compare PAT-based pitch tracking with the well-known pitch tracking tool ESPS Get_f0 (Talkin 1995) for the following four measures. The voiced error (VE) denotes the percentage of voiced frames misclassified as unvoiced, the unvoiced error (UE) is defined as the inverse case, the gross pitch error (GPE) denotes the percentage of voiced frames at which the estimation and the reference pitch frequency differ by more than 20%, and the root mean squared (RMS) difference (in Hertz) is computed between the estimated and reference pitch frequencies when there are no gross pitch errors. The Get_f0 result is taken from Sha (2004).

Table 1 shows the implementation configuration of PAT. For PAT, we perform unsupervised training on each sentence. After training, Viterbi decoding is performed to determine the most likely pitch contour. During decoding, we set the frames with energy below 1% of the sentence's average energy to be silence frames, and there is no constraint on the frames whose energy is above this threshold. The result is given in Table 2.

As can be seen from Table 2, while PAT performs comparable with Get_f0 in U/V decision, it performs much more accurately in estimating the pitch frequency. To further confirm this conclusion, we conduct another experiment to eliminate the possible effect of U/V decision on the pitch estimation accuracy. We perform pitch tracking only on the labeled voiced frames that are also correctly classified by Get_f0 as voiced. In this case, only the GPE and RMS measures are relevant. The results are given in Table 3. For clear comparison, we also report the 95% confidence interval for the mean squared errors of pitch estimates using the bootstrap method (Bisani and Ney 2004) separately for males and females. The superiority of PAT over get_f0 is obvious.

3.2 Speech synthesis

In this experiment, the usefulness of PAT for parametric speech synthesis is tested. After we perform unsupervised training of PAT on each Edinburgh sentence, we can synthesize each sentence using the estimated pitch frequency and vocal tract response. A formal subjective listening test is conducted to evaluate the naturalness and clearness of the synthesized speech.

We use the overlap-add (OLA) method to synthesize the speech waveform from the reconstructed magnitude spectrogram using the phase of the original speech (Quatieri, 2001). There are two possible reconstruction of the magnitude spectrogram based on PAT, which are given by the following formula respectively:

$$x_t^{Z\_SYNTHESIS} = \Gamma_{\hat{l}_t} \cdot E[z_t | x_{1:T}] \tag{26}$$

$$x_t^{MU\_SYNTHESIS} = E[a_t | x_{1:T}] \cdot \Gamma_{\hat{l}_t} \mu_{\hat{c}_t} \tag{27}$$

where $\hat{c}_t$ and $\hat{l}_t$ are the most likely phonetic cluster and pitch frequency determined by the Viterbi decoding based on the unsupervised trained PAT model, and $E[\cdot]$ denotes the expectation. Both reconstructions use the inferred pitch frequencies. The difference is that the first reconstruction (called *Z_SYNTHESIS*) uses the inferred z_t, while the second reconstruction (called *MU_SYNTHESIS*) uses the mean vocal tract shape according to the inferred phonetic cluster c_t.

The LPC-based speech synthesis is conducted for comparison. In this case, the U/V decision and the pitch frequencies are provided using the results from Get_f0. For voiced region, the LPC-based spectral envelope is used for synthesis, while the unvoiced region is directly taken from the original speech. For fair comparison, the LPC order is set to be 60, which is equal to the number of DCT coefficients used in PAT to model the spectral envelope.

Table 1: Implementation configuration of PAT for pitch tracking

Sampling frequency (Hz)	8000
Frame length (ms)	30
Frame shift (ms)	10
FFT size N	240
Number of DCT coefficients M	60
EM iterations	3
Discretized pitch frequency range (midi)	35 - 67.4
Discretized frequency interval (midi)	0.15
Number of discretized pitches L	217
Number of voiced clusters K_v	15
Number of unvoiced/silence clusters K_u	15
α in Equ. (6) to constrain U/V transition	20

Table 2: Pitch tracking results on Edinburgh database

	PAT	Get_f0
UE (%)	5.38	8.84
VE (%)	4.83	4.29
GPE (%)	0.91	2.86
RMS (Hz)	5.46	5.83

Table 3: Pitch tracking results with U/V labeling

	PAT	Get_f0
GPE (%)	1.51	2.07
RMS (Hz)	5.4556	5.7792
95% confidence interval for the mean squared errors of pitch estimates		
Male	(10.89, 20.57)	(12.65, 21.78)
Female	(46.22, 78.74)	(60.65, 97.22)

Table 4: MOS score comparison in the speech synthesis experiments on the Edinburgh sentences

	PAT	others
Z_SYNTHESIS vs. LPC	4.33	2.21
Z_SYNTHESIS vs. original	4.37	4.69
MU_SYNTHESIS vs. LPC	3.24	2.31
MU_SYNTHESIS vs. original	3.34	4.98

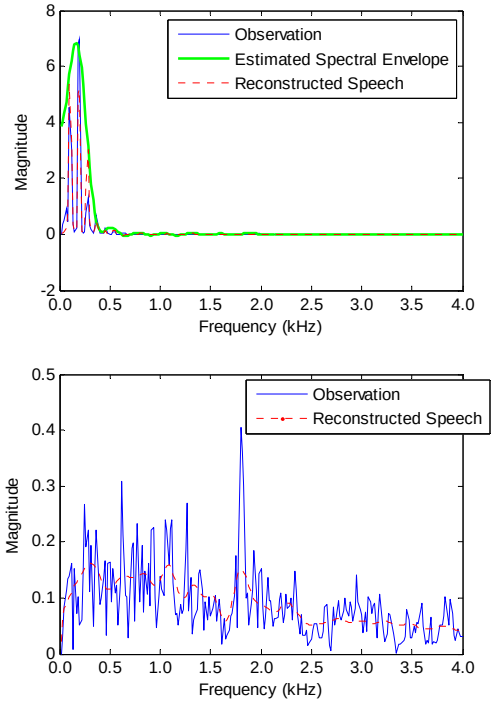


Figure 4: Examples of voiced (top) and unvoiced (bottom) speech and the corresponding *Z_SYNTHESIS*-based reconstructed speech.

Four comparison experiments are conducted. Both *Z_SYNTHESIS*-based and *MU_SYNTHESIS*-based synthesized speech are compared with LPC-based synthesized speech and the original speech. Five listeners with normal hearing participate in the experiment. Each of them listens to 80 different pairs of sentences, with 20 pairs for each comparison. For a pair of tested sentences, the listener is asked to grade each sentence with a MOS score. The MOS (mean opinion score) is expressed as a single number in the range 1 to 5, where 1 is lowest perceived speech quality, and 5 is the highest perceived speech quality.

It can be seen from the MOS grading result in Table 4 that both *Z_SYNTHESIS* and *MU_SYNTHESIS* outperform LPC-based synthesis. The quality of *Z_SYNTHESIS*-based reconstructed speech is close to that of the original speech. Some examples of *Z_SYNTHESIS*-based reconstructed speech are shown in Figure 4. Although the quality of *MU_SYNTHESIS* is not as good as that of *Z_SYNTHESIS*, it has the advantage of using a smaller number of parameters to code the speech. It can recover the magnitude spectrum of each frame using 1 double ($E(a_t|x_t)$) and 2 integers ($\hat{c}_t$ and $\hat{l}_t$), and has the potential for low-rate speech coding.

3.3 Phoneme clustering

In this experiment, we test the capability of PAT to learn meaningful structure from unlabeled speech. As we know, the vocal tract response is known to determine the phoneme being uttered, while the presence of pitch and

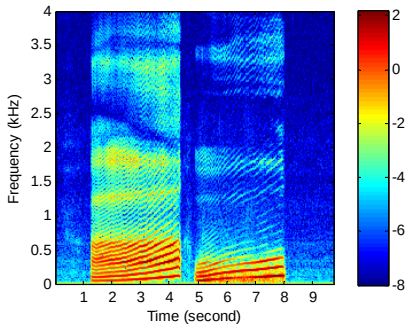


Figure 5: Speech spectrogram for unsupervised phoneme clustering. The former part corresponds to the phoneme /**a:**/, while the later is the phoneme /**u:**/. There is a clear rising of pitch frequencies for each phoneme.

the pitch frequency mainly determines the prosodic aspect of speech such as the stress, rhythm, and intonation. If we perform clustering (a typical unsupervised learning technique) on the speech frames, it would be more useful for different clusters to represent different phonemes, instead of mixed representation of phonetic and prosodic information.

In the PAT model, the vocal tract response and the pitch are separately modeled due to their different roles in the production of speech. Thus, it has the potential to learn phoneme clusters, independent of the pitch frequencies. In contrast, if we perform clustering directly on the MFCCs (mel-frequency cepstral coefficients) based on HMMs, the clustering results will be less meaningful. This is due to the fact that the MFCC feature roughly estimates the vocal tract shape by summing the outputs from the triangular filter banks equally spaced on the mel-frequency scale, and is sensitive to pitch changes, as shown in Stephenson (2004).

To verify the above analysis, an utterance is recorded by pronouncing /**ɑ:**/ and /**u:**/ with a rising tone while holding the vocal tract shape. As can be seen in Figure 5, the pitch frequencies change significantly, while the formants remain stable. We train a PAT containing K_v=4 voiced clusters and K_u =3 unvoiced/silence clusters on the utterance shown in Figure 5. Figure 6(a) shows the clustering result from performing Viterbi decoding using the trained PAT. Although we initialize 4 voiced clusters, the trained PAT successfully learns two non-trivial voiced clusters and assigns the voiced frames correctly to the corresponding clusters. For comparison, we also trained a HMM containing 7 states/clusters on MFCCs of the same utterance. Figure 6(b) gives the clustering result based on the trained HMM, which clearly shows its failure to discover the meaningful phoneme clusters.

3.4 Pitch tracking under noisy conditions

Theoretically, we can employ factorial HMM modeling approach to perform pitch tracking under noisy conditions. In this paper, we use a simple method to test PAT-based pitch tracking under additive white Gaussian noises. In this case, the expected magnitude spectrum of the white

Table 5: PAT-based Pitch tracking under noisy conditions

	10 dB	5 dB	0 dB
GPE (%)	0	1.58	7.74
RMS (Hz)	1.3052	2.2190	6.2605

noise is constant, denoted as σ_n. We estimate σ_n from the first several frames of the noisy utterance that are assumed to be speech-free. The distribution of the noise term n_t is modified as follows to accommodate the external noise:

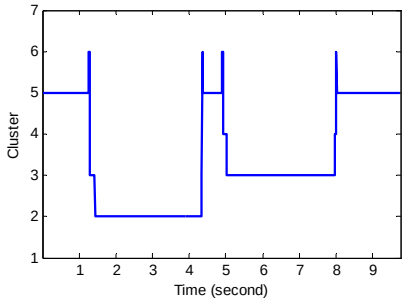
$$p(n_t|c_t = c) = \mathcal{N}\left(n_t; \sigma_n, m_c^2(\Psi + \mathrm{DIAG}[\sigma_n^2])\right) \quad (28)$$

where $\mathrm{DIAG}[\sigma_n^2]$ denotes the diagonal matrix whose diagonal corresponds to the vector σ_n^2. In the experiment, we record two utterances from two male speakers. Each speaker utters the five vowels - /**a:**/, /**ɔ:**/, /**ɜ:**/, /**i:**/, /**u:**/ in one utterance. We first perform supervised training of a PAT containing K_v =5 voiced clusters and no unvoiced/silence clusters on the utterance from speaker 1, given the beginning and the end of each vowel and constraining frames in each vowel region to fall in the specified voiced cluster. After the modification as shown in Equ. (28), the trained PAT is used for pitch tracking with U/V labels on the noisy utterance from speaker 2, corrupted by white Gaussian noise with different SNRs. For the test utterance, the pitch tracking on the clean speech as described in Section 3.1 is performed and the result is used as the ground truth. The resulting GPE and RMS measures are given in Table 5, which clearly shows the effectiveness of PAT-based pitch tracking on the noisy utterance.

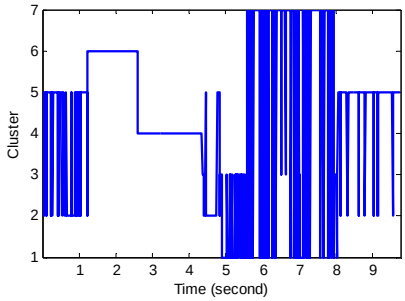
Furthermore, we can achieve speech enhancement by using the *Z_SYNTHESIS* method as described in Section 3.2 to reconstruct the clean utterance, based on the inferred z_t and l_t from the noisy utterance. It can be seen from Figure 7 that the enhanced speech is very close to the original speech by observing the spectrogram. For speech enhancement, it is widely known that most signal filtering methods, e.g. spectral subtraction and wiener filtering, suffer from some residual noise known as musical noise (Quatieri, 2001). We can hardly hear any musical noise in the PAT-based enhanced speech, as provided in the supplementary material.

4 Conclusions

Most speech analysis/synthesis systems are based on the basic physical model of speech production - the acoustic tube model. Examining the drawbacks with current speech analysis methods motivate us to directly model the spectrogram and to integrate together the three fundamental speech parameters - the pitch, energy and spectral envelope. We successfully devise such a model called probabilistic acoustic tube (PAT) model. The integration is performed in a principled manner with explicit physical meaning. Once we have such a probabilistic generative model of speech, a variety of speech analysis/synthesis tasks can be reduced to inference and learning in this model. We demonstrate the capability



(a)



(b)

Figure 6: The clustering result from performing Viterbi decoding: (a) using the trained PAT, where cluster 1-4 are voiced clusters, (b) using the trained HMM on MFCCs.

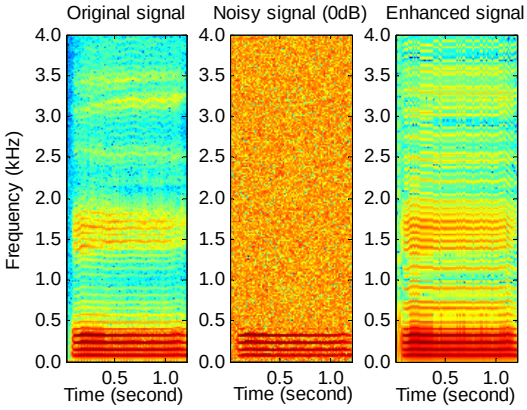


Figure 7: Speech enhancement result for the vowel /**ɔ:**/.

of PAT for a number of speech analysis/synthesis tasks, such as pitch tracking under both clean and additive noise conditions, speech synthesis, and phoneme clustering. It can be easily seen that PAT could be further applied for computational acoustic scene analysis and noise-robust speech recognition which are our future woks.

Acknowledgements

This work is supported by National Natural Science Foundation of China (61075020).

References

T.F. Quatieri, *Discrete-time speech signal processing: principles and practice*, Prentice Hall, 2001.

J. Li, et al., A unified framework of HMM adaptation with joint compensation of additive and convolutive distortions, *Computer Speech and Language*, no. 3, vol. 23, 2009.

F.R. Bach, M.I. Jordan, Discriminative training of hidden Markov models for multiple pitch tracking, *Proc. ICASSP*, 2005.

H. Kameoka, et al., A multipitch analyzer based on harmonic temporal structured clustering, *IEEE Trans. Audio, Speech and Language Processing*, 2006.

T.A. Stephenson, M.M. Doss, H. Bourlard, Speech recognition with auxiliary information, *IEEE Trans. Audio, Speech and Language Processing*, 2004.

B.J. Frey, N. Jojic, Estimating mixture models of images and inferring spatial transformation using the EM algorithm, *Proc. CVPR*, 1999.

M. Reyes-Gomez, N. Jojic, D. Ellis, Deformable spectrograms, *Proc. AISTATS*, 2005.

J.R. Hershey, et al., Super-human multi-talker speech recognition: a graphical modeling approach, *Computer Speech and Language*, 2010.

D. Talkin, *Speech coding and synthesis*, Elsevier Science, 1995.

P.C. Bagshaw, S.M. Hiller, M.A. Jack, Enhanced pitch tracking and the processing of f0 contours for computer aided intonation teaching, *Proc. Eurospeech*, 1993.

F. Sha, J.A. Burgoyne, L.K. Saul, Multiband statistical learning for f0 estimation in speech, *Proc. ICASSP*, 2004.

M. Bisani, H. Ney, Bootstrap estimates for confidence intervals in ASR performance evaluation, *Proc. ICASSP*, 2004.