Vocal Tract Spectrogram Estimation with Formant Frequency Contour Factorization*

Yunhan Zou[1,2], Li Li[1,3], Hirokazu Kameoka[1]
[1]NTT Communication Science Laboratories
[2]Georgia Institute of Technology [3]University of Tsukuba

1 Introduction

The spectral envelope of speech contains rich information about the voice characteristics of the speaker. According to the source-filter model for speech production, spectral envelopes correspond to the resonance characteristics of the vocal tract. If we can assume that the vocal tract spectra observed at each time frame can be approximated as a linear sum of spectral templates scaled by time-varying amplitudes, the vocal tract spectrogram, interpreted as a non-negative matrix, can be approximated as the product of two non-negative matrices, one containing the time-independent spectral templates arranged as column vectors and the other containing the time-varying amplitudes arranged as row vectors. This way of representing a data matrix is called non-negative matrix factorization (NMF) and applying it to a vocal tract spectrogram has several attractive features.

First, it allows us to decouple vocal tract spectra into time-independent and time-dependent factors, namely the spectral templates and the temporal activations. This decomposition is noteworthy since the former factor roughly describes the voice characteristics of the speaker whereas the latter contains information about the transcription of the uttered sentence. Thus, if we convert the spectral templates while keeping the temporal activations unchanged, we can modify a speaker's voice to sound like it were spoken by another speaker [1, 2]. This technique is called voice conversion.

It also provides a novel solution for tackling the source-filter decomposition problem. The source-filter model assumes that a speech spectrum is given as the product of the vocal tract and the glottal excitation spectra. Source-filter decomposition refers to the problem of estimating the vocal tract and the excitation source spectra solely from a speech spectrum. Within a voiced segment where we assume the excitation source signal to be a periodic pulse train, the spectrum of the produced speech is given by the product of the vocal tract spectrum and the equally spaced pulses with an interval equal to the fundamental frequency (F_0). This process is shown in Fig. 1. Thus, a speech spectrum can be seen as a sampled version of the vocal tract spectrum with missing information between the harmonics. Widely used vocoders such as STRAIGHT [3] and WORLD [4] provide ways of extracting a spectral envelope from a voiced spectrum by smoothly interpolating between the harmonics. However, it

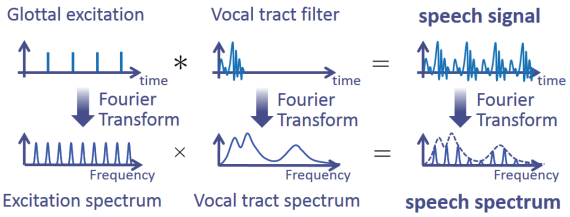

Fig. 1 Source-filter model

is generally impossible to impute the missing data and restore the true vocal tract spectrum only from the speech spectrum observed at a particular frame. This illustrates the limitation of the frame-by-frame analyzers that perform source-filter decomposition at each time frame independently. An attempt has recently been made to overcome this limitation by leveraging multi-frame observations. The idea is to employ a missing data imputation technique using NMF to estimate the entire vocal tract spectrogram [5]. The underlying assumption behind this method is that the vocal tract spectra are expected to be represented fairly well as a linear sum of a limited number of spectral templates. This expectation may be supported by the fact that the number of vowel types is usually limited in normal speech.

As these examples show, factorization of vocal tract spectrogram has significant potential. In the previous work, NMF provides a template-based representation of spectral magnitude. However, spectral magnitude is not the only important factor in characterizing speech. As shown in Fig. 2, vocal tract spectrograms are typically characterized by several dominant peaks varying continuously over time. These peaks are often called formants. The frequencies of these peaks correspond to the resonance frequencies of the vocal tract and the frequencies of the first two formants are known to be important in determining the quality of vowels. Fig. 3 shows an example of the spectrogram of Fig. 2 as a result when NMF is applied to the decomposition of spectral magnitude only. By looking at the flat formant trajectories in this example, we can confirm that the NMF model is not well suited to express the continuous trajectories of the formant frequencies. This is because the NMF model is only able to apply a linear sum of spectral templates in expressing a spectrum and does not have the ability to interpolate the peak frequencies of the templates. To obtain a template-based representation better suited to express vocal tract spectrogram, we propose yet another model consisting of formant frequency set templates.

*フォルマント周波数行列分解による声道スペクトログラム推定, 鄒 雲漢 (NTT, ジョージア工科大), 李 莉 (NTT, 筑波大), 亀岡 弘和 (NTT)

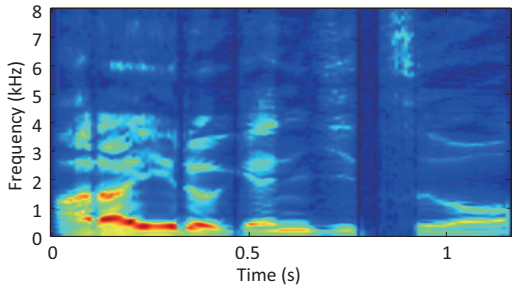

Fig. 2 Vocal tract spectrogram.

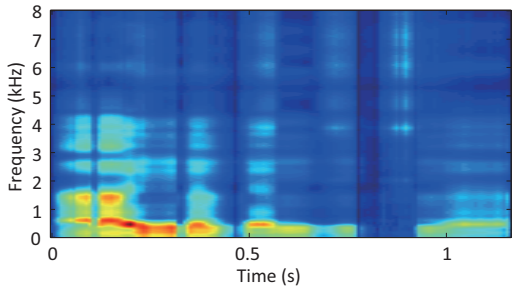

Fig. 3 NMF applied to the spectrogram of Fig. 2.

2 Model

2.1 Motivation

Each type of vowels is characterized by a formant frequency set. During the transition from a phoneme to another, one formant frequency set varies continuously towards another formant frequency set. Since the phoneme number is limited, it would be reasonable to assume that the formant frequency set at each time frame can be represented as a linear combination of formant frequency set templates. This idea implies that a matrix that contains the formant frequency values at different time frame, which we hereafter call "formant frequency matrix", can be modeled as the product of two matrices. In the following, we propose a novel vocal tract spectrogram model by incorporating this formant frequency matrix model. We also derive a convergence-guaranteed algorithm for estimating the parameters from an observed spectrogram.

2.2 Formant frequency matrix factorization

We start by describing a spectral envelope using a Gaussian mixture model (GMM) [6, 7], interpreted as a function of frequency (see Fig. 4 for its graphical illustration):

$$F(\omega, t) = \beta(t) \sum_{i=1}^{I} \alpha_i(t) G_i(\omega, t), \tag{1}$$

$$G_i(\omega, t) = \frac{1}{\sqrt{2\pi}\sigma_i(t)} e^{-\frac{(\omega - \mu_i(t))^2}{2\sigma_i^2(t)}}, \tag{2}$$

where ω and t denote frequency and time, respectively, and I is the number of Gaussian mixture components. This representation consists of parameters (roughly) corresponding to the frequency and power of spectral peaks and is particularly convenient for incorporating the above idea. $\mu_i(t)$, $\sigma_i^2(t)$ and $\alpha_i(t)$

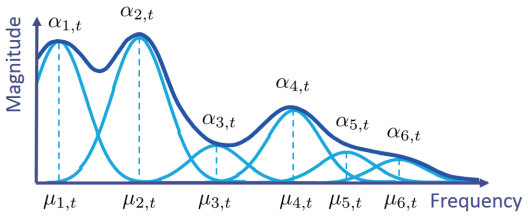

Fig. 4 GMM representation for spectral envelope.

are the mean, variance and weight of each Gaussian component, corresponding to the frequency, peakiness and power of a dominant peak in the spectral envelope. To eliminate an indeterminacy in the scalings of $\beta(t)$ and $\alpha_i(t)$, we assume $\alpha_i(t)$ to satisfy a sum-to-one constraint so that $\beta(t)$ represents the energy of the spectrum in each frame.

If we assume that each Gaussian corresponds to a formant, $\mu_i(t)$ can be seen as the frequency of the i-th formant at time t. As discussed in 2.1, we assume $\mu_i(t)$ to be represented as a weighted sum of K formant frequency set templates $w_{i,1}, \ldots, w_{i,K}$. K can thus be thought of as the number of vowel types. In the same way, we assume the magnitude $a_{i,k}$ to be represented as a weighted sum of K magnitude set templates $a_{i,1}, \ldots, a_{i,K}$. Here, we constrain the weight coefficients $h_k(t)$ for $w_{i,k}$ and $a_{i,k}$ to be equal:

$$\mu_i(t) = \sum_k w_{i,k} h_k(t), \tag{3}$$

$$\alpha_i(t) = \sum_k a_{i,k} h_k(t). \tag{4}$$

The way these parameters are tied to each other is extremely important since we may want to associate a $\{w_{i,k}, a_{i,k}\}$ pair with a particular shape of the i-th prototype spectral envelope. Similarly, for the variance we assume

$$\sigma_i^2(t) = \sum_k c_{i,k} d_k(t). \tag{5}$$

Since formant frequencies usually vary continuously over time, we may want $h_k(t)$ to be a smooth function. One convenient way of ensuring smoothness is to express $h_i(t)$ as a weighted sum of half-overlapping Hanning functions $g_1(t), \ldots, g_J(t)$ that cover the entire time range:

$$h_k(t) = \sum_j g_j(t) l_{j,k}, \tag{6}$$

and treat $l_{j,k}$ as the parameter to estimate instead of $h_k(t)$. In this way, the smoothness of $h_k(t)$ can be easily controlled by window size settings.

We place the first Gaussian function to center at $\omega = 0$ to account for the large energy present at lower frequencies in the vocal tract spectrogram. Namely, we set $w_{1,k} = 0$.

3 Parameter estimation

We use Θ to denote the set of the parameters:

$$W = \{w_{i,k}\}_{i,k}, \tag{7}$$

$$A = \{a_{i,k}\}_{i,k}, \tag{8}$$
$$L = \{l_{j,k}\}_{j,k}, \tag{9}$$
$$C = \{c_{i,k}\}_{i,k}, \tag{10}$$
$$D = \{d_k(t)\}_{k,t}, \tag{11}$$
$$\beta = \{\beta(t)\}_t. \tag{12}$$

Note that all these values must be non-negative. Here we derive a parameter estimation algorithm that fits the present model $F(\omega, t)$ to an observed spectrogram $Y(\omega, t)$. For convenience of the derivation of the optimization algorithm, we use the Kullback-Leibler (KL) divergence (also known as the I divergence) to measure the difference between the present model and an observed spectrogram:

$$I(\Theta) = \iint \Bigg(Y(\omega,t) \log \frac{Y(\omega,t)}{F(\omega,t)} - Y(\omega,t) + F(\omega,t) \Bigg) d\omega dt. \tag{13}$$

Note that the integral taken over the interval $[0, \infty)$ of the first Gaussian centered at $\omega = 0$ is $1/2$ and that of each of the other Gaussians is approximately equal to 1. Thus, to make them consistent, we redefine $G_1(\omega, t)$ as $\frac{2}{\sqrt{2\pi}\sigma_1(t)} \exp\{-\frac{(\omega-\mu_1(t))^2}{2\sigma_1^2(t)}\}$. Since the Hanning functions are set to satisfy $\sum_j g_j(t) = 1$, we can show that the integral of $F(\omega, t)$ taken over the interval $[0, \infty)$ is $\beta(t)$ when

$$\sum_{i=1}^{I} a_{i,k} = 1, \tag{14}$$
$$\sum_{k} l_{j,k} = 1, \tag{15}$$

are satisfied. (13) can therefore be rewritten as

$$I(\Theta) = \iint \left(Y(\omega,t) \log \frac{Y(\omega,t)}{F(\omega,t)} - Y(\omega,t) \right) d\omega dt + \int \beta(t) dt, \tag{16}$$

under the redefined $G_1(\omega, t)$ and the conditions (14) and (15).

In addition, since we want each Gaussian to be associated with a dominant peak in a spectrum, we do not want the magnitude of any of the Gaussians to be extremely small. Thus, to prevent the magnitude matrix $a_{i,k}$ from becoming sparse in the process of model fitting, we incorporate a penalty term given by the negative logarithm of a symmetric Dirichlet distribution. A symmetric Dirichlet distribution is maximized when all the arguments become exactly equal. The penalty term is thus given as

$$R(A) = -\log \left(\prod_{i,k} {a_{i,k}}^{\varphi} \right). \tag{17}$$

Hence, our objective function to be minimized is

$$J(\Theta) = I(\Theta) + R(A). \tag{18}$$

Although it is difficult to analytically find the global minimum point of the current optimization problem, we can search for a stationary point using a majorization-minimization (MM) algorithm. An MM algorithm refers to an iterative algorithm that consists of iteratively minimizing an auxiliary function called a "majorizer". Suppose $\mathcal{J}(\Theta)$ is an objective function that we wish to minimize with respect to Θ. A majorizer $\mathcal{J}^+(\Theta, \Xi)$ is then defined as a function satisfying $\mathcal{J}(\Theta) = \min_{\Xi} \mathcal{J}^+(\Theta, \Xi)$, where Ξ is called an auxiliary variable. An algorithm that consists of iteratively minimizing $\mathcal{J}^+(\Theta, \Xi)$ with respect to Θ and Ξ is guaranteed to converge to a stationary point of the objective function. For our objective function, we can show that

$$\begin{aligned} J^+(\Theta,\Xi) = & \iint Y(\omega,t) \Bigg[\log Y(\omega,t) - \log \beta(t) \\ & - \sum_i \lambda_i(\omega,t) \Bigg\{ \sum_{k,j} \theta_{i,k,j}(t) \log \frac{a_{i,k} g_j(t) l_{j,k}}{\theta_{i,k,j}(t)} \\ & - \frac{1}{2} \left(\frac{\sum_k c_{i,k} d_k(t) - \zeta_i(t)}{\zeta_i(t)} + \log \zeta_i(t) \right) \\ & - \frac{1}{2} \left(\sum_k \frac{\rho_{i,k}^2(t)}{c_{i,k} d_k(t)} \right) \times \\ & \left(\omega^2 - 2\omega \sum_{k,j} w_{i,k} g_j(t) l_{j,k} + \sum_{k,j} \frac{w_{i,k}^2 g_j^2(t) l_{j,k}^2}{\gamma_{i,k,j}(t)} \right) \\ & + \log \lambda_i(\omega,t) \Bigg\} \Bigg] d\omega dt + \eta \\ & + \int \beta(t) dt + R(A), \end{aligned} \tag{19}$$

is an auxiliary function where Ξ is the set of the auxiliary variables

$$\lambda = \{\lambda_i(\omega,t)\}_{i,\omega,t}, \tag{20}$$
$$\theta = \{\theta_{i,k,j}(t)\}_{i,k,j,t}, \tag{21}$$
$$\zeta = \{\zeta_i(t)\}_{i,t}, \tag{22}$$
$$\rho = \{\rho_{i,k}(t)\}_{i,k,t}, \tag{23}$$
$$\gamma = \{\gamma_{i,k,j}(t)\}_{i,k,j,t}, \tag{24}$$

that must satisfy $0 \le \lambda_i(\omega, t) \le 1$, $\sum_i \lambda_i(\omega, t) = 1$, $0 \le \theta_{i,k,j}(t) \le 1$, $\sum_{k,j} \theta_{i,k,j}(t) = 1$, $0 \le \rho_{i,k}(t) \le 1$, $\sum_k \rho_{i,k}(t) = 1$, $0 \le \gamma_{i,k,j}(t) \le 1$, and $\sum_{k,j} \gamma_{i,k,j}(t) = 1$. The update equation for each of the parameters and the auxiliary variables can be obtained in closed form through (19). The details are omitted owing to space limitations.

4 Experiments

4.1 Experimental Conditions

To show the ability of the proposed method to express a vocal tract spectrogram with natural formant trajectories using a compact template-based representation, we tested the proposed method on the spectrogram obtained with the STRAIGHT analysis.

The STRAIGHT analysis was executed with 5 [ms] time shift and the sampling frequency of the test signal was 16 [kHz]. In the experiment, we used $K = 10$ templates in our model. $w_{i,k}$ describing the i-th formant location in the k-th template was initialized by observing the formant locations from the STRAIGHT spectrogram. In addition, it was initialized to include $I = 5$ Gaussians. The Hanning function $g_j(t)$ was designed to half-overlap with each covering approximately 100 [ms] time span. We used speech samples from the ATR speech database. The parameter update process is divided into a total of 5 parts: during the initial fitting, basis matrix $w_{i,k}$ is unchanged while other parameters are updated for 70 iterations. This is done under the expectation that with other parameters initialized to proper values, the subsequent update on $w_{i,k}$, the key to this experiment, will therefore be more efficient and less error-prone. Then follows the part where formant location set $w_{i,k}$ joins the update along with all others for 120 iterations. In the last 3 parts, the number of Gaussians is further doubled (9, 17 and 33) to make the fitting more accurate and each part runs for 120 iterations.

4.2 Experimental Results

To confirm the model fitting accuracy, we plotted the spectrograms of specific segments obtained with the proposed method along with the estimated formant trajectories $\mu_i(t)$. These graphs can be found in Figs. 5 and 6. We can see from these graphs that the estimated formant trajectory is not over-smoothed and follows the formant transition well.

To verify the reliability of extracted formant locations, columns of matrix $w_{i,k}$ and $a_{i,k}$ were swapped in a certain way to obtain a new speech signal. By listening to this speech signal, the speech was converted in an utterance with a completely different sentence while it still sounded natural enough as if it was spoken by the same speaker. Modified spectrogram is plotted to show the same segments as in the previous step, as in Fig. 7.

As the final step of evaluation, the distance between first and second formant locations in $w_{i,k}$ is expanded. By listening to the generated speech, we concluded that speech element sounds more distinguished from each other. The modified spectrogram can be seen in Fig. 8.

5 Conclusion

To obtain a template-based representation well suited to express vocal tract spectrograms, this paper proposed a vocal spectrogram model that allows us to co-factorize the formant frequency/magnitude matrices in the process of model fitting. We developed an update algorithm for the model parameters based on an auxiliary function approach. We confirmed through experiments that the formant contour factorization is able to represent speech formant trajectory, which resulted in natural-sounding synthetic speech.

Acknowledgment This work was supported by JSPS KAKENHI Grant Numbers 26730100 and 26280060.

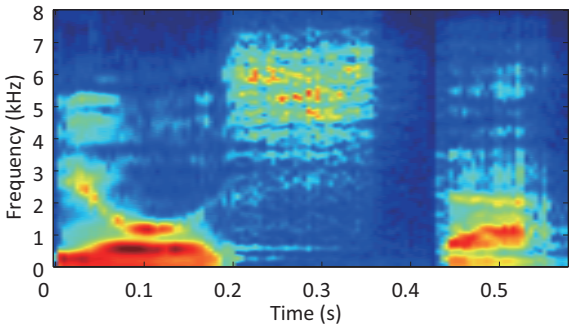

Fig. 5 STRAIGHT spectrogram $Y(\omega, t)$.

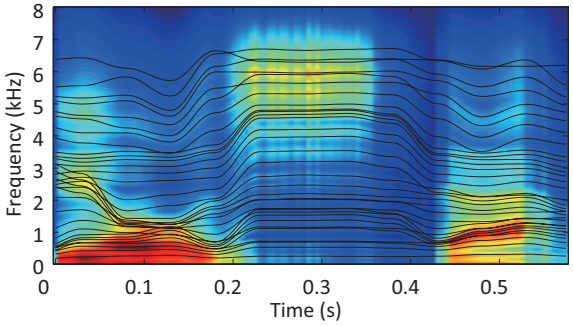

Fig. 6 $F(\omega, t)$ fitted to $Y(\omega, t)$ along with $\mu_i(t)$.

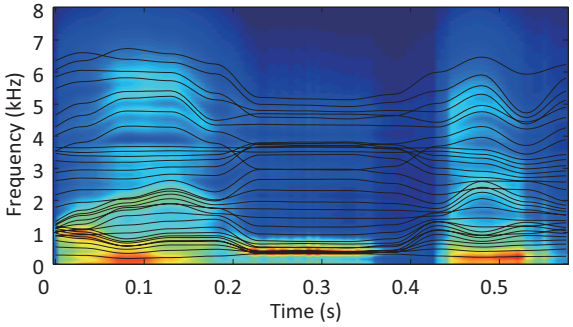

Fig. 7 Reconstructed $F(\omega, t)$ obtained by swapping columns of W and A.

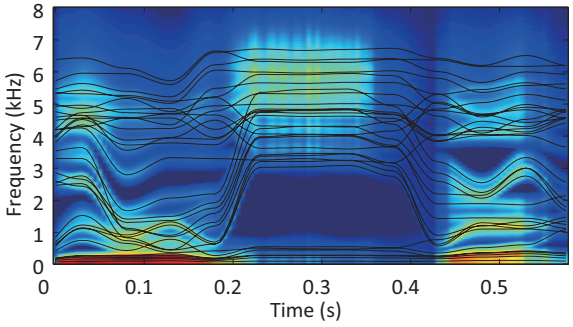

Fig. 8 Reconstructed $F(\omega, t)$ with modified W.

参考文献

[1] Takashima et al., IEICE Trans. IS, E96-A(10), 1946–1953, 2013.
[2] Wu et al., IEEE/ACM Trans. ASLP, 22(10), 1506–1521, 2014.
[3] Kawahara et al., Proc. ICASSP, 3933–3936, 2008.
[4] Morise et al., IEICE Trans. IS, E99-D(7), 1877–1884, 2016.
[5] 中村, 亀岡, 音講論 (春), 3-P-33, 393–396, 2016.
[6] Kameoka et al., IEEE Trans. ASLP, 18(6), 1507–1516, 2010.
[7] Hojo et al., Proc. SSW8, 129–134, 2013.