HIDDEN SPECTRAL PEAK TRAJECTORY MODEL FOR PHONE CLASSIFICATION

Yiu-Pong LAI, Man-Hung SIU

Hong Kong University of Science and Technology
Department of Electrical and Electronic Engineering
Clear WaterBay, Hong Kong
harry@ust.hk, eemsiu@ust.hk

ABSTRACT

It is well known that spectrogram readers can classify different phones from their spectral-time characteristics, such as the formants. In this paper we present a novel acoustic model for phone classification based on the implicit estimation of the spectral peak trajectory as a polynomial time function. By making use of the known relationship between the spectral peak information and the cepstral coefficients, cepstral-based phone trajectories are built as functions of the hidden spectral trajectories. This captures the intuitive formant trajectories in the spectral domain while allowing speech modeling to be done in the more familiar cepstral domain. We have evaluated this hidden spectral peak trajectory model in both vowel classification and phone classification tasks. On a simple single Gaussian model, the hidden spectral peak trajectory model outperforms the HMM on both vowel and phone classification tasks. The new can also be combined with the HMM model. This combination performs better than a more complex HMM with similar number of parameters.

1. INTRODUCTION

Segmental models have been proposed to capture speech temporal correlations explicitly so as to relax the conditional independence assumptions in HMM [1]. For example, the polynomial trajectory model, which represents the temporal correlation as a polynomial time function, has been shown to work well for both vowel and phone classification [2] [3]. However, when the time function is defined in the cepstral domain, which is the typical domain for speech recognition, it is hard to visualize the shape because of the high dimensionality. If the first and second order cepstral coefficients are also modeled using polynomial trajectories, there is also a consistency problem since they should be, by itself the time derivatives of the cepstral coefficients.

The spectrogram, on the other hand, can easily be visualized and human experts can distinguish between different phones by reading their time-spectral patterns, in particular, the formant patterns. Formants are the time function of resonances frequencies of the vocal tract, which corresponds to the tracks of high energy points (spectral peaks) in the spectrogram. Because of the distinguishing features of the formant patterns, the formants and formants related parameters have been used for vowel classification [4]. Furthermore, the formants can be sufficiently represented by second or third order polynomials.

While it is desirable to model the time correlation using formants or related features that can be visualized and sufficiently represented by a low order polynomial, it is also desirable to keep the speech recognition modeling in the cepstral domain. Some reseaches suggested that a mapping function to be used to connected these two domains [5]. This can also be achieved by taking advantage of the relationship between the poles and zeros of the complex spectrum and the cepstral coefficients. Applying trajectory model in the spectral domain has the added advantage that a consistent set of cepstral and cepstral derivative time functions can be derived.

In this paper we propose a hidden spectral peak polynomial trajectory model (HSPTM) by assuming speech segments to be generated by a set of hidden spectral peak trajectories. Each peak is defined by two variables, the peak location and the bandwidth which are equivalent to the phase and magnitude of a pole in the complex spectrum. The trajectories of these peaks are assumed to be polynomial time functions. Using the relationship between spectral peaks and cepstral, similar to the LPC analysis, the spectral trajectories in turn define a set of time varying cepstral trajectories that are modeled as the means of the segment. By making the spectral peak trajectory hidden, one avoids the difficult problem of explicitly estimating the spectral track.

The rest of the paper is organized as follows. In Section 2, the spectral peaks representation for speech is presented. In Section 3, the hidden modeling technique for the spectral peak trajectory is introduced including the estimation of parameters and likelihood computation. We also describe the prior models and phone a duration models, both of which improve the classification accuracy. In Section 4, the experiments for both vowel classification and phone classification are presented. The work is summarized in Section 5.

2. SPECTRAL PEAKS REPRESENTATION OF SPEECH

Instead of representing the formant tracks, which are not defined for unvoiced sounds, we focus on modeling the poles of the complex spectrum similar to the approach taken in linear prediction (LP) analysis [6]. It can be shown that the center frequencies and the bandwidths of the formants can be computed from the roots of the predictor polynomial when an LP is used [7]. Furthermore, for a given set of poles, the cepstral coefficients can be computed.

Consider the predictor polynomial 2p poles,

$$\begin{aligned} A(z) &= 1 - \sum_{i=1}^{2p} a_i z^{-i} \\ &= \prod_{i=1}^{2p} (1 - z_i z^{-1}). \end{aligned} \tag{1}$$

Let H(z) be the spectrum of the speech frame. Using the all-pole model,

0-7803-8484-9/04/$20.00 ©2004 IEEE

$$\begin{aligned} H(z) &= \frac{G}{A(Z)} \\ &= \frac{G}{\prod_{i=1}^{p}(1-z_i z^{-1})(1-z_i^* z^{-1})} \end{aligned} \quad (2)$$

Note that the roots z_i can be expressed as

$$z_i = e^{-\pi \frac{b_i}{f_s} + j2\pi \frac{f_i}{f_s}}, \quad (3)$$

where f_s is the sampling rate, b_i, f_i are the bandwidth, the center frequency of the i-th root respectively.

Taking the logarithm on the transfer function and using the Taylor series expansion, we have

$$\begin{aligned} &\log H(z) \\ &= \log(G) - \sum_{i=1}^{p} \log(1 - z_i z^{-1}) - \sum_{i=1}^{p} \log(1 - z_i^* z^{-1}) \\ &= \log(G) + \sum_{i=1}^{p}\sum_{k=1}^{\infty} \frac{z_i^k z^{-k}}{k} + \sum_{i=1}^{p}\sum_{k=1}^{\infty} \frac{z_i^{*k} z^{-k}}{k} \\ &= \log(G) + \sum_{k=1}^{\infty}(\sum_{i=1}^{p} \frac{1}{k} e^{-\pi k \frac{b_i}{f_s}} (e^{j2\pi k \frac{f_i}{f_s}} + e^{-j2\pi k \frac{f_i}{f_s}}))z^{-k} \\ &= \log(G) + \sum_{k=1}^{\infty}(\sum_{i=1}^{p} \frac{2}{k} e^{-\pi k \frac{b_i}{f_s}} \cos(2\pi k \frac{f_i}{f_s}))z^{-k} \end{aligned} \quad (4)$$

Given the roots, the cepstral coefficients can be computed by taking inverse z-transforms.

$$c_k = \sum_{i=1}^{p} \frac{2}{k} e^{-\pi k \frac{b_i}{f_s}} \cos(2\pi k \frac{f_i}{f_s}), \quad \text{for } k > 0. \quad (5)$$

c_k is often called the LPC cepstrum because of the use of the all-pole assumption for the spectrum.

In speech recognition, the derivatives of the cepstral coefficients are part of the features. Using Equation 5, the cepstral derivatives can also be expressed as a function of the hidden spectral peaks. To simply the discussion the rest of the theoretical development assumes that only the cepstrum are used as features with the understanding that it can easily be extended to include cepstral derivatives.

3. SPECTRAL PEAK TRAJECTORY MODEL

One key piece of information in formants is the time correlation of the track such as its shape and slope. These can not be captured in a single short-time speech frame. Similarly for the spectral peaks, a time function is needed to represent the temporal information. Polynomial trajectory models have been proposed in for modeling cepstral temporal function with some success [2] and can also be applied to capture spectral-time correlation and the resulting track be mapped back to the cepstral domain.

Applying the parametric trajectory modeling on the spectral peaks, both the magnitude $B_i(n)$ and $F_i(n)$ are assumed to be generated by the polynomial functions on a normalized time scale as described in [2]. For an N-frame long speech segment, $B_i(n)$ and $F_i(n)$ are given by

$$\begin{aligned} \hat{F}_i(t) &= \omega_{i,0} + \omega_{i,1}t + \omega_{i,2}t^2 + \cdots + \omega_{i,d-1}t^{d-1}, \\ \hat{B}_i(t) &= \beta_{i,0} + \beta_{i,1}t + \beta_{i,2}t^2 + \cdots + \beta_{i,d-1}t^{d-1}, \end{aligned}$$

where $\omega_{i,t}$ and $\beta_{i,t}$ are the polynomial coefficients for the i-th track and $t = \frac{n-1}{N-1}$.

3.1. Modeling speech segments in cepstral domain

In the cepstral domain, the speech frames are modeled in the same fashion as in other segment model. That is, frame $x(n)$ of an N-frame segment X_1^N is viewed as generated by a time-varying trajectory $\mu_\phi(n)$ and a zero-mean residue $e_\phi(n)$ [2]. That is,

$$x(n) = \mu(n) + e(n). \quad (6)$$

What is different is how $\mu_\phi(n)$ is modeled. Denote the complex spectrum of $x(n)$ as $H(n, z)$. Equations 2-5 can be modified by adding the time index n. Thus, the i-th bandwidth and i-th spectral peak, k-th cepstrum become $b_i(n)$, $f_i(n)$ and $c_k(n)$ respectively. Equation 5 is rewritten as,

$$\begin{aligned} c_k(n) &= \sum_{i=1}^{p} \frac{2}{k} e^{-\pi k \frac{b_i(n)}{f_s}} \cos(2\pi k \frac{f_i(n)}{f_s}), \text{for } k > 0 \\ &= \sum_{i=1}^{p} \frac{2}{k} e^{-\pi k B_i(n)} \cos(\pi k F_i(n)), \text{for } k > 0 \end{aligned} \quad (7)$$

where $B_i(n) = b_i(n)/f_s$, is a function of the magnitude of the spectral peak, and $F_i(n) = 2f_i(n)/f_s$ is the normalized spectral peak frequency in Hz.

The likelihood of the observation against model ϕ can be written as,

$$\begin{aligned} \mathcal{L}_\phi(x_1^N) &= \mathcal{L}(x_1^N; \beta_\phi, \omega_\phi, \Sigma_\phi) \\ &= \sum_{n=1}^{N}[-\frac{K}{2}\log(2\pi) - \frac{1}{2}\log(|\Sigma_\phi|) \\ &\quad -\frac{1}{2}(x(n) - \mu_\phi(n))^T \Sigma_\phi^{-1}(x(n) - \mu_\phi(n))], \end{aligned} \quad (8)$$

where Σ_ϕ is the variance of ϕ and K is the dimension of $x(n)$.

Denote $\mu_\phi(n, k)$ as the k-th component of $\mu_\phi(n)$.

$$\mu_\phi(n, k) = \sum_{i=1}^{p} \frac{2}{k} e^{-\pi k B_{i,\phi}(n)} \cos(\pi k F_{i,\phi}(n)),$$

where F_ϕ, B_ϕ are the trajectories for the peak frequency and the magnitude trajectory of phone ϕ respectively.

In the rest of the paper, ϕ will be removed from the variable to simplify our notation unless it is needed to differentiate between different phone models.

3.2. Parameter Estimation

The parameters in the hidden spectral peak trajectory models include the ω_i's, β_i's and the Σ's for each phonetic model. These parameters can be estimated by maximizing the observation likelihood defined in Equation 8. Because a closed form solution cannot be obtained, the gradient descent method is used. The updated parameters $\hat{\omega}$ and $\hat{\beta}$ are given by:

$$\begin{aligned} \hat{\omega}_{i,j} &= \omega_{i,j} - \epsilon \frac{\partial \mathcal{L}(x_1^N)}{\partial \omega_{i,j}} \\ &= \omega_{i,j} - \epsilon \sum_{n=1}^{N} \frac{-\partial \mu(n)^T}{\partial \omega_{i,j}} \Sigma^{-1}(x(n) - \mu(n)) \end{aligned}$$

where

$$\frac{\partial \mu(n, k)}{\partial \omega_{i,j}} = -2\pi e^{-\pi k B_i(n)} \sin(\pi k F_i(n)) \left(\frac{n-1}{N-1}\right)^j, \quad (9)$$

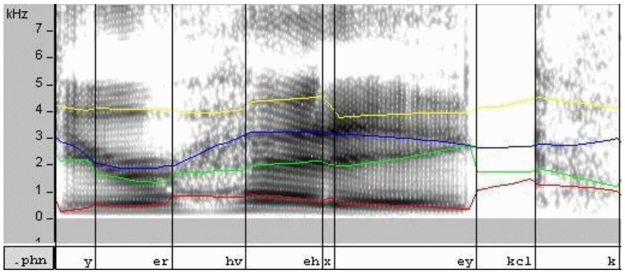

Fig. 1. Spectral peaks of an utterance ‘your headache’.

and

$$\begin{aligned}\hat{\beta}_{i,j} &= \beta_{i,j} - \epsilon \frac{\partial \mathcal{L}(x_1^N)}{\partial \beta_{i,j}} \\ &= \beta_{i,j} - \epsilon \sum_{n=1}^{N} \frac{-\partial \mu(n)^T}{\partial \beta_{i,j}} \Sigma^{-1}(x(n) - \mu(n)),\end{aligned}$$

where

$$\frac{\partial \mu(n,k)}{\partial \beta_{i,j}} = -2\pi e^{-\pi k B_i(n)} \cos(\pi k F_i(n)) \left(\frac{n-1}{N-1}\right)^j. \quad (10)$$

To maintain meaningful spectral peak tracks, constraints are imposed on the locations of the tracks to restrict their range not to exceed the sampling frequency, and keeping the spectral tracks non-crossing. In Figure 1, the spectrogram of the utterance ‘your headache’ and the corresponding estimated hidden spectral peak tracks, is shown. It shows that the estimated quadratic spectral peak trajectory can closely follow the visual formants.

3.3. Time-varying Variance

[8] proposes that time varying variance across a phone motivated by the fact that co-articulation effect could have more influence on phone boundaries. They show that phone classification accuracy can be improved with the time varying variance. In their work, a piece-wise constant variance function is used in which a segment is uniformly partitioned into multiple region. Each region has its own variance. Similar approach can be applied to HSPSM. If a phone is partitioned into M regions $\zeta = \{\zeta_1, \ldots, \zeta_M\}$ of constant variance Σ_s, $1 \leq s \leq M$, these variances can be estimated simply by

$$\Sigma_s = \frac{\sum_{n \in \zeta_s} (x(n) - \mu(n))^T (x(n) - \mu(n))}{Count(N_s)}, \quad (11)$$

where $Count(N_s)$ is the number of frames belonging to segment s.

The log-likelihood of the segment becomes

$$\begin{aligned}\mathcal{L}(x_1^N) &= \sum_{s=1}^{M} \sum_{n \in \zeta_s} [-\frac{K}{2}\log(2\pi) - \frac{1}{2}\log(|\Sigma_s|) \\ &\quad -\frac{1}{2}(x(n) - \mu(n))^T \Sigma_s^{-1} (x(n) - \mu(n))]. \quad (12)\end{aligned}$$

3.4. Prior and Duration Probability

In general, different phones occurs at unequal frequency. A prior probability can be used to improve the performance of any phone classifier. This prior probability can be estimated by counting the number of occurrences in the training corpus. The duration model is also introduced to improve the classification performance as there are great differences in the duration of different phones. The duration probability $p(N|\phi)$ is modeled, following a Gamma distribution, and the parameters of the distribution are estimated from the training utterances.

With the prior probability and the duration probability, the classification criteria can be written as,

$$\hat{\phi} = \arg\max_{\phi} [\mathcal{L}_\phi(x_1^N) + w_d \log(p(N|\phi)) + w_p \log(p(\phi))],$$

where w_p and w_d are the weighting for the prior probability and duration probability as the dynamic range of the log-likelihood is significantly larger.

3.5. Combination with HMM phone classifier

Besides using the formants frequencies trajectory model, we also evaluated the combined HMM and formant frequencies trajectory model by means of a weighted combination of their log-likelihoods.

$$\begin{aligned}\hat{\phi} &= \arg\max_{\phi} [(1-\alpha)\mathcal{L}_\phi(x_1^N) + w_d \log(p(N|\phi)) \\ &\quad + w_p \log(p(\phi)) + \alpha \tilde{\mathcal{L}}_\phi(x_1^N)],\end{aligned}$$

where $\tilde{\mathcal{L}}_\phi(x_1^N)$ HMM model likelihood and α is a constant weighting on the two models.

4. EXPERIMENTS

To evaluate the performance of hidden spectral peak trajectory model, a set of experiments on a speaker independent vowel classification and phone classification using the TIMIT database were performed.

For both classification tasks, only ’sx’ and ’si’ sentences were used for the training and testing. There are a total of 3696 training utterances and 1344 testing utterances and the standard TIMIT training and test sets were used. The acoustic features consisted of the first 12 LPCC with 12 poles (excluding the zeroth coefficients) and their first order derivative. The utterance based Cepstral Mean Subtraction (CMS) was applied to remove the channel effect. For comparison, a three-state HMM model was trained for each phone. For the spectral peak trajectory model, six spectral peaks were used to model the spectral-time characteristics and the variance of residual was assumed to be diagonal. While typically three to four formants are sufficient to identify vowels, six peaks are used partly to be consistent with the 12 LPC poles used in frontend processing and partly to allow us more resolution as well as ability to capture the shape of consonants. In addition, the effect of using prior probability (P), duration probability (D), and the combined HMM and spectral peak trajectory model, were evaluated.

4.1. Vowel Classification

The vowel classification task consisted of 16 different vowels and diphthongs, /aa, ae, ah, ao, aw, ay, er, eh, ey, iy, ih, ow, oy, uh, uw, ux/. The classifier was trained using the alignments and the 61 TIMIT labels defined in the TIMIT database. The results are summarized in Table 1.

These results show that prior probability and duration probability play an important role in classification. Furthermore, the

Model	# parameters per model	Accuracy
Spectral Peaks	108	47.18%
Spectral Peaks + P	109	50.08%
Spectral Peaks + D	111	50.56%
Spectral Peaks + P + D	112	52.60%
HMM + P	145	48.60%
HMM (2 mixtures) + P	289	50.81%
Spectral Peaks combined with HMM + P + D	256	53.21%

Table 1. *Performance of Vowel Classification; P: prior probability, D: duration probability.*

Model	# parameters per model	Accuracy
Spectral Peaks	108	47.10%
Spectral Peaks + P	109	51.73%
Spectral Peaks + D	111	50.87%
Spectral Peaks + P + D	112	55.77%
HMM + P	145	54.70%
HMM (2 mixtures) + P	289	57.20%
Spectral Peaks combined with HMM + P + D	256	58.06%

Table 2. *Performance of Phone Classification; P: prior probability, D: duration probability.*

hidden spectral peak trajectory model performs significantly better than HMM. While the acoustic models currently tested are simply single Gaussian models, the positive results are quite encouraging. The combination of the HMM and the hidden spectral peak trajectory model is better than using a HMM with two mixtures shows that the hidden peak trajectory model is capturing useful information not captured by the extra mixture in HMM.

4.2. Phone Classification

In the task of phone classification, 48 models were training according to the phone list from Lee [9], and the classification results were folded into 39 different phone classes to determine the classification accuracy. The results are tabulated in Table 2.

The results for the phone classification are similar to those obtained for the vowel classification. This indicates that consonants can also be classified using the spectral peak trajectory although their formants do not exist. The combined HMM and the spectral peak trajectory model shows improvements in the accuracy comparing to both single mixtures or two mixtures HMM.

5. DISCUSSION AND CONCLUSION

In this paper we proposed a novel approach for modeling speech. We proposed a new cepstral trajectory model which, exploiting the relationship between the spectral peaks and cepstral features, is derived from the hidden center frequencies and the bandwidth of the spectral peaks that are modeled as polynomial time functions. For the voiced phones, the center frequencies of the spectral peaks represent the formants. While there are no formants in unvoiced speech sounds, experimental results, as well as experience from LPC analysis, showed that the set of spectral peaks can sufficiently represent the unvoiced sounds.

The preliminary results reported in this paper for this new model are very encouraging. While the performance of the hidden spectral peak trajectory model itself is no better than that of the HMM, several limitations, such as the constraint on the frame alignment within the model, the small number of mixtures, can be relaxed. Furthermore, new improvements in segmental models can also be integrated. Finally, the combined HMM and trajectory model ensures the best results.

6. REFERENCES

[1] M. Ostendorf, V. V. Digalakis, and O. A. Kimball, "From hmm's to segment models: A unified view of stochastic modeling for speech recognition," vol. 4, pp. 360–387, 1996.

[2] H. Gish and K. Ng, "A segmental speech model with applications to word spotting," in *Proceedings of ICASSP 93*, 1993, vol. 2, pp. 447–450.

[3] W. J. Holmes and M. J. Russell, "Probabilistic-trajectory segmental hmms," vol. 13, pp. 3–37, 1999.

[4] P. Schmid and E. Barnard, "Explicit, n-best formant features for vowel classification," in *Proceedings of ICASSP 97*, 1997, vol. 2, pp. 21–24.

[5] M. J. Russell and P. J. B. Jackson, "The effect of intermediate articulatory layer on the performance of a segmental hmm," in *Proceedings of Eurospeech 03*, 2003, pp. 2737–3740.

[6] L. Rabiner and B. H. Juang, *Fundamentals of speech recognition*, Prentice Hall Inc, 1993.

[7] X. D. Huang, A. Acero, and H. W. Hon, *Spoken language processing: A guide to theory, algorithm and system development*, Prentice Hall Inc, Upper Saddle River, New Jersey, 2000.

[8] H. Gish and K. Ng, "Parametric trajectory models for speech recognition," in *Proceeding of ICSLP 96*, 1996, vol. 1, pp. 466–469.

[9] K. F. Lee and H. W. Hon, "Speaker-independent phone recognition using hidden markov models," vol. 37, pp. 1641–1648, 1989.