STOP CONSONANT CLASSIFICATION BY DYNAMIC FORMANT TRAJECTORY

Yanli Zheng , Mark Hasegawa-Johnson and Sarah Borys

ECE Department, University of Illinois at Urbana-Champaign
{zheng3, jhasegaw, sborys}@uiuc.edu

ABSTRACT

LPC analysis is one of the most powerful techniques in speech analysis. Spectral zeros during consonant or consonant-vowel transition regions introduce difficulties in estimating LPC parameters. In this paper, we propose to estimate formant frequencies from LPC model by MUSIC (Multiple Signal Classification) and ESPRIT (Estimation of Signal Parameters via Rotational Invariance Techniques). Formant candidates estimated by LS (Least Square), MUSIC and ESPRIT are combined to find an optimal solution. The effectiveness of this algorithm is verified by place classification task of stop consonants.

1. OVERVIEW

Classification of stop consonants remains one of the most challenging problems in speech recognition. Halberstadt (1998) [3] reported classification of phones in the TIMIT database using heterogeneous acoustic measurements, they found that: for vowel classification, listener-labeler error and machine error produce very similar performance; for place classification of stop consonants, machine classification results lag by a factor of $1.8 - 5.1$. Sussman (1991) [9] investigated the locus equation and applied it to the place classification of stop consonants.They found that discriminant analysis using $F2_{onset}$ and $F2_{vowel}$ as predictors showed 76% classification accuracy. And they achieved 100% classification accuracy using derived slope and intercept values as predictors. It is generally agreed that relative invariance cues of stop consonant place are coded in dynamic spectral shape starting from stop release. Sussman's result suggested that a compact representation of dynamic spectra could be found by accurate formant estimation.

Formant frequencies are estimated from the LPC model. Spectral zeros during consonant or consonant-vowel transition regions introduce difficulties in estimating LPC parameters. This paper proposes an algorithm to improve formant estimation by combining formant candidates from different estimators. The rest of the paper is organized as follows: Section 2 reviews important properties of the LPC model; in Section 3, MUSIC and ESPRIT are proposed for formant estimation; in Section 4, an algorithm combining formant estimation of LS, MUSIC and ESPRIT is proposed, and the effectiveness of this algorithm is demonstrated by place classification of stop consonants.

This work was supported by NSF award number 0132900. Statements in this paper reflect the opinions and conclusions of the authors, and are not endorsed by the NSF.

2. LPC MODEL

Discrete-time speech production model can be described by [5]:

$$Y(z) = G(z)T(z)R(z) \tag{1}$$

where $G(z)$ is z-transform of source, $R(z)$ is the radiation impedance, and $T(z)$ is the transfer function of the vocal tract taking the form of an ARMA model.

In the time-domain, for stationary process $\{y_t\}$ with $E[y_t] = 0$, the ARMA model of Eq. (1) is:

$$\sum_{k=0}^{p} a_k y_{n-k} = \sum_{m=0}^{q} b_m w_{n-m}, \quad \text{p, q are even} \tag{2}$$

where $p = 2M$ and $\{w_t\}$ are i.i.d. $N(0, \sigma_w^2)$ white noise. For ARMA model in Eq.(2),the autocorrelation ρ_k satisfies the normal equation [11] :

$$\sum_{j=0}^{p} a_j \rho_{k-j} = 0, \quad k \geq (q+1) \tag{3}$$

For a finite-length data record, the sample autocorrelation function of $\hat{\rho}_k$ is used. The homogeneous solution of Eq. (3) for $\hat{\rho}_k$ will be superimposed exponential signals in noise:

$$\hat{\rho}_k = r_k + v_k = \frac{1}{2}\sum_{m=1}^{p} \eta_m \lambda_m^k + v_k = \sum_{m=1}^{M} \eta_m e^{-k\sigma_m} cos(k\omega_m) + v_k \tag{4}$$

$$\lambda_m = \begin{cases} e^{-\sigma_m - j\omega_m}, & m = 1, 2, \cdots, M \\ e^{-\sigma_m + j\omega_m}, & m = M+1, \cdots, 2M \end{cases} \tag{5}$$

where $M = p/2$ is the number of formant frequencies, v_k is modeled as Gaussian noise to represent modeling and measuring error. Eq. (4) can be expressed in the vector form

$$\vec{\rho} = C(\omega)\vec{\eta} + \vec{v} \tag{6}$$

where

$\vec{\rho}$ $N \times 1$ noise-corrupted autocorrelation sample vector;

$C(\omega) = [\vec{b}_1, \cdots, \vec{b}_m, \cdots, \vec{b}_M]$ $N \times M$ matrix, with cosine bases for $\vec{\rho}$;

$\vec{\eta}$ $M \times 1$ amplitude vector of cosine basis;

$\vec{v}$ $N \times 1$ Gaussian noise vector;

$\vec{b}_m = [1, e^{-\sigma_m}\cos(\omega_m), \cdots, e^{-(N-1)\sigma_m}\cos((N-1)\omega_m)]^T$.

Axiom 1. *Given Eq. (6), assuming that $\vec{v} \sim N(0, \sigma_v^2 I_N)$, Maximum Likelihood estimation of η and σ_v^2 are as follows:*

$$\hat{\vec{\eta}} = (C^T C)^{-1} C^T \vec{\rho} \tag{7}$$

$$\hat{\sigma}_v^2 = \frac{\vec{\rho}^T Q_\omega \vec{\rho}}{N - m} \tag{8}$$

where $Q_\omega = I_N - C(C^T C)^{-1} C^T$, and Q_ω is also a projection matrix (i.e. $Q^2 = Q$, $Q = Q^T$).

When σ_m in Eq. (5) is small, columns of C will be approximately orthogonal to each other.

Lemma 1. *For matrix $C(\omega) = [\vec{b}_1, \cdots, \vec{b}_m, \cdots, \vec{b}_M]$, if*

$$\frac{\vec{b}_k^T \vec{b}_j}{|\vec{b}_k||\vec{b}_j|} = \begin{cases} 1 & k = j \\ 0 & k \neq j \end{cases}$$

then

$$\hat{\eta}_m = \frac{b_m^T \vec{\rho}}{b_m^T b_m} = \frac{b_m^T \vec{\rho}}{||b_m||^2} \tag{9}$$

$$\hat{\vec{\rho}}_m = \hat{\eta}_m \vec{b}_m \tag{10}$$

$$||\hat{\rho}_m||^2 = \frac{||b^T \rho||^2}{||b_m||^2} \tag{11}$$

$$\hat{\sigma}_v^2 = \frac{||\rho||^2 - \sum_{m=1}^{M} ||\hat{\rho}_m||^2}{N - M} \tag{12}$$

Theorem 1. *[1] Let N samples of $\rho(k)$, generated according to a p^{th} order model as in Eq.(3), form the $(N-L)\times(L+1)$ Toeplitz matrix*

$$A_f = \begin{bmatrix} \rho(L) & \rho(L-1) & \cdots & \rho(0) \\ \rho(L+1) & \rho(L) & \cdots & \rho(1) \\ \vdots & \vdots & & \vdots \\ \rho(N-1) & \rho(N-2) & \cdots & \rho(N-1-L) \end{bmatrix} \tag{13}$$

Then $rank(A_f) = min(p, L, N-L)$.

Proof.

$$A_f = \sum_{k=1}^{p} \eta_k \begin{bmatrix} \lambda_k^L & \cdots & 1 \\ \lambda_k^{L+1} & \cdots & \lambda_k \\ \vdots & & \vdots \\ \lambda_k^{N-1} & \cdots & \lambda_k^{N-1-L} \end{bmatrix} = G(\lambda) \Lambda H(\lambda)^T \tag{14}$$

where

$$\Lambda = \begin{bmatrix} \eta_1 & 0 & \cdots & 0 \\ 0 & \eta_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & \eta_p \end{bmatrix}$$

$G(\lambda) = [\mathbf{g}(\lambda_1), \cdots, \mathbf{g}(\lambda_p)]$, $\mathbf{g}(\lambda) = [1, \lambda, \lambda^2, \cdots, \lambda^{N-1-L}]^T$ and

$H(\lambda) = [\mathbf{h}(\lambda_1), \cdots, \mathbf{h}(\lambda_p)]$, $\mathbf{h}(\lambda) = [\lambda^L, \lambda^{L-1}, \cdots, 1]^T$

Since both H and G are Vandermonde matrix, when λ_k's are different, $rank(A_f) = min(p, L, N-L)$. □

Define Y, X and W to be

$$Y = X + W = A_f^T + W = HS + W \tag{15}$$

where

$$S = \Lambda G^T = [\mathbf{s}_1, \cdots, \mathbf{s}_p]^T$$

and W is a noise matrix.

3. PARAMETER ESTIMATION

In order to estimate p complex eigenvalues, $L > p$ and $N > p + L$ are used for data matrix A_f. Rao [7] proved that increasing filter length L decreases the sensitivity of the estimated parameters to noise and reduces the numerical ill conditioning. As $N \to \infty$, the sample correlation matrix $\hat{\Sigma}$ converges to Σ,

$$\hat{\Sigma} \to \Sigma = HE[SS^H]H^H + \sigma_n^2 I = H\Sigma_s H^H + \sigma_n^2 I \tag{16}$$

$$\Sigma = [U_s\ U_n] \begin{bmatrix} \sigma_1^2 + \sigma_n^2 & & & & & & \\ & \ddots & & & & & \\ & & \sigma_p^2 + \sigma_n^2 & & & & \\ & & & \sigma_n^2 & & & \\ & & & & \ddots & & \\ & & & & & \sigma_n^2 \end{bmatrix} \begin{bmatrix} U_s^H \\ U_n^H \end{bmatrix} \tag{17}$$

where $[U_s, U_n]$ are the eigenvectors of the sample correlation matrix $\hat{\Sigma}$, and U_s spans the signal space, and U_n spans the noise space, and they are also left singular vectors of data matrix Y. Eq. (16) shows that asymptotically the effect of the white noise is simply to shift the eigenvalues of Σ_s by σ_n^2.

3.1. MUSIC Solution

It is clear in Eq.(15) that H has full column rank, and S has full row rank. In the absence of noise, $dim(Y) = dim(X) = p$. Eq.(14) shows that $R(X) \subset R(H)$, therefore $R(X) = R(H)$ (i.e. range of X is the same as range of H). Since $R(U_s) = R(X) = R(H) = R^\perp(U_n)$, we have

$$U_n^H H = 0 \tag{18}$$

λ_k's can be found by Root-MUSIC algorithm [6].

3.2. ESPRIT Solution

Y in Eq. (15) can be partitioned into two sets that can be expressed as:

$$Y_1 = X_1 + W_1 = H_1 S + W_1, \quad Y_2 = X_2 + W_2 = H_1 \Phi S + W_2 \tag{19}$$

where Y_1 is formed by deleting the top row out of Y, Y_2 is formed by deleting the last row out of Y, $\Phi = diag(\lambda_1, \lambda_2, \cdots, \lambda_p)$. Φ can be solved by ESPRIT algorithm in [10]. Formant frequencies can be estimated from phase of λ_k's.

4. EXPERIMENT

This section illustrates an improved formant tracking method by combining formant estimations from LS (e.g. Lattice solution or Durbin's recursive solution), MUSIC and ESPRIT. The effectiveness of this method will be shown in a classification task of stop place. Stop consonants /b,d,g,p,t,k/ in this experiment are extracted from TIMIT database. Training tokens are from TIMIT TRAIN, testing tokens are from TIMIT TEST. Number of tokens used in the experiment are shown in Table 1.

Table 2. Prior Knowledge of the Formant Frequencies (F_m) and their Bandwidth(B_m)

Formant	F_1	F_2	F_3	F_4	F_5	F_6
Range (Hz)	[200 1000]	[800 3200]	[1200 3900]	[3000 6000]	[3500 7000]	[4500 7000]
Bandwidth (Hz)	70	70	70	100	100	100
	min $(F_m - F_{m-1})$					
m	1	2	3	4	5	6
Distance (Hz)	-	120	120	300	300	300

Table 1. Number of tokens used in the experiment

Spoken Place	Labial		Alveolar		Velar	
	b	p	d	t	g	k
# of TRAIN tokens	2181	2588	2432	3948	1191	3794
# of TEST tokens	886	975	841	1367	452	1204

4.1. Experiment Setup

4.1.1. Prior Knowledge of Formant Parameters

Locations of the formant frequencies on the frequency axis are constrained by the physical limitations of the vocal tract. In our experiment, formant frequencies F1 to F6 are estimated. The frequency ranges and shortest distance between adjacent formants are listed in Table 2. To pay more attention to the local maximum in the spectrum, we also use fixed bandwidth from Table 2 in our calculation.

4.1.2. System Design

ASR systems in [2] and [12] used [F1, F2, F3] for the task of digit recognition. We choose to track six formants (F1 to F6) in our system because formant parameters in the range of F4 to F6 provide important discriminant information for sound classification. A two-channel filter-bank system is used for formant tracking. The system flow chart is shown in Figure 1. A 15ms analysis window and 5 ms frame rate are used. $p = 12$ (i.e. 6 formant candidates in each subband), $L = 16$, $N = 29(= L + p + 1)$.The cutoff frequency of LP filter is 3500 Hz. The cutoff frequency for HP filter is 3000 Hz. (TIMIT database is sampled at 16 KHz). For each data frame $\vec{y}_t$, an optimal set of formant candidates $\vec{f}^* = [F_1\ F_2\ F_3\ F_4\ F_5\ F_6]^T$ is found.

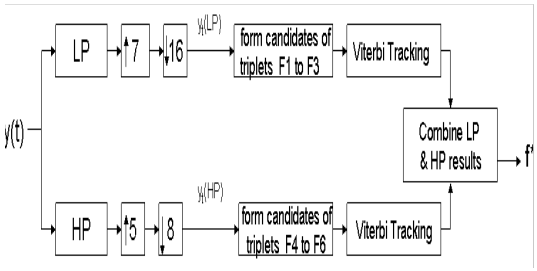

Fig. 1. System Block

The procedure of forming candidates of the triplet of [F1 F2 F3] is as follows:

1. Pass y_t through a LP filter and downsample it to get $y_t(LP)$;
2. Find the sample autocorrelation function coefficients $\hat{\rho}_k$ ($k = 0, 1, \cdots, N-1$);
3. Estimate candidates of formants by LS, Root-MUSIC and ESPRIT;
4. If N formant candidates are found in the above step, form $\binom{N}{3}$ triplets, then keep N^* triplets which satisfy constraints in Table 2;
5. For each qualified triplet, $\hat{\sigma}_v^2$ is found by Eq.(8) or Eq. (12);

Similarily, triplets of F4 to F6 are formed from $y_t(HP)$. Then Viterbi tracking [13] is used to find the optimal formant sequence.

4.2. Test Result

4.2.1. Static Spectrum at Stop Release

Figure 2 illustrates the attributes of burst spectra. The data are averages of automatic measurements from syllables consisting of /b,d,g,p,t,k/ followed by each of 18 different vowels and semivowels (two tokens of each C-V syllable is extracted from TIMIT, one from a female talker, and one from a male talker).

Figure 2 shows that:

- Amplitudes of labial and alveolar do not differ in the F2 and F3 range;
- In the F4 and F5 region, the curves of labial and alveolar begin to diverge, and at F5 the amplitude of /t,d/ is about 6 dB greater than that of the /b,p/ burst;
- Velar stop has a relatively high amplitude in the range of F2 to F4.

This measurement is consistent with the analysis in [8].

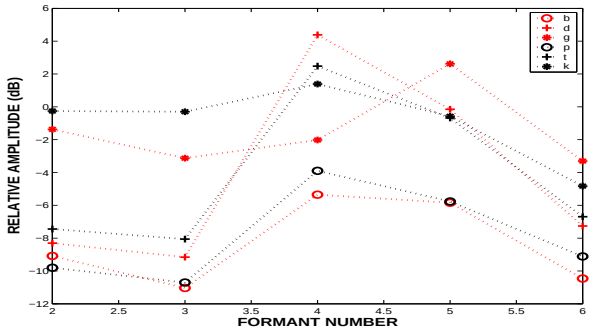

Fig. 2. Average difference (in decibels) of the spectrum amplitude of the burst for labial(o), alveolar(+) and velar(*) stops.

4.2.2. Dynamic Spectrum Analysis

Automatic measurements of the F1 movement following the release are compared in Figure 3 for labial, alveolar and velar stop consonants, with the following vowel /ae/ by a male talker from TIMIT. Right after the consonant release, measurements of F1 cannot be found very accurately, therefore by assuming that the starting frequency of F1 at the instant of release is 200 Hz, Figure 3 shows that the F1 transition is fastest for the labial and slowest for the velar. The result is consistent with Stevens' calculation [8], which presumably reflects the different rates of movement and different lengths of the constriction.

Five Frames (35 ms) are used in the classification task, which is started at the stop release. The five vectors are concatenated to a long vector for classification. A 20-dimensional vector of formant parameters includes F1 to F6, η_1 to η_6, $||\hat{\rho}_1||^2$ to $||\hat{\rho}_6||^2$, and energies from both LP and HP filters. A 36-dimensional new feature is formed by concatenating MFCC and formant parameters. Figures 4 and 5 are LDA analyses of MFCC and the new feature. LDA plot of the new feature shows a better class-separation than MFCC. Classification result of stop place is given in Table 3. SVM [4] classifier with RBF kernel function is used. The new feature with formant parameters consistently outperforms MFCC parameters.

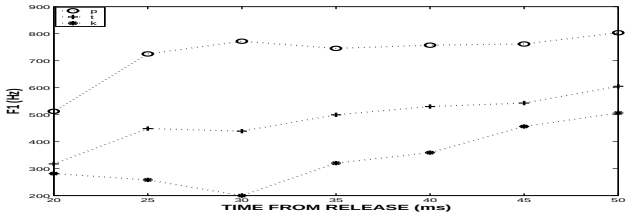

Fig. 3. Measurement of the first-formant frequency

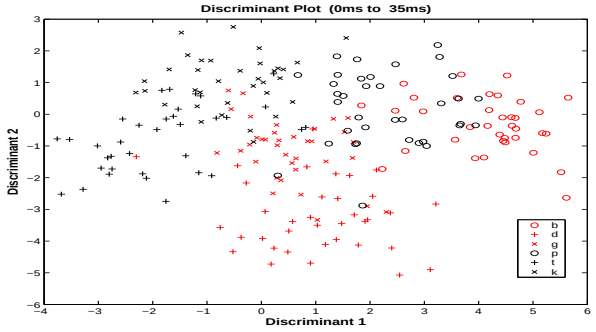

Fig. 4. Discriminant Analysis of MFCC (35ms duration). For each of the stop consonants [b,d,g,p,t,k], 36 tokens are used in this plot.

Table 3. SVM classification of stop place of TIMIT TEST

	Formant Parameters + MFCC	MFCC + Delta + Accel.
Labial	95.0%	85.1%
Alveolar	87.2%	87.1%
Velar	85.1%	82.9%

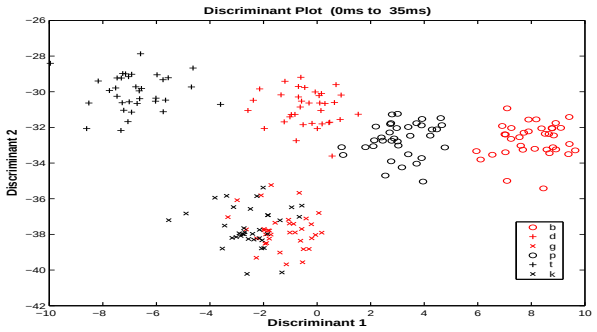

Fig. 5. Discriminant Analysis of New Feature (35ms duration). For each of the stop consonants [b,d,g,p,t,k], 36 tokens are used in this plot.

5. REFERENCES

[1] Yoram Bresler, Samit Basu, and Christophe Couvreur. *Hilbert Space and Least Squares Methods for Signal Processing*. Unpublished lecture notes, 2000.

[2] Marcia A. Bush and Gary E. Kopec. Network-based connected digit recognition. *IEEE Trans. Acoustics, Speech, and Signal Processing*, ASSP-35(10):1401–1413, October 1987.

[3] Andrew K. Halberstadt. *Heterogeneous Acoustic Measurements and Multiple Classifiers for Speech Recognition*. PhD thesis, MIT, Cambridge, MA, Nov. 1998.

[4] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference and Prediction*. Springer, 2001.

[5] T. F. Quatieri. *Discrete-Time Speech Signal Processing*. Prentice-Hall PTR, NJ, 2001.

[6] Bhaskar Rao. Performance analysis of root-music. In *IEEE Trans. Acoustics, Speech, and Signal Processing*, pages 1939–1949, 1989.

[7] D. V. N. Rao. Perturbation analysis of a SVD based method for the harmonic retrieval problem. In *Proc. ICASSP*, pages 624–627, 1985.

[8] Kenneth N. Stevens. *Acoustic Phonetics*. MIT Press, Cambridge, MA, 1999.

[9] Harvey M. Sussman, Helen A. McCaffrey, and Sandra A. Matthews. An investigation of locus equations as a source of relational invariance for stop place categorization. *J. Acoust. Soc. Am*, 90(3):1309–1325, September 1991.

[10] Henry L. van Trees. *Optimum array processing: Part IV of Detection, Estimation and Modulation Theory*. Wiley-Interscience, 2002.

[11] William Wei. *Time Series Analysis: Univariate and Multivariate Methods*. Addison-Wesley Publishing Company, 1994.

[12] Lutz Welling and Hermann Ney. Formant estimation for speech recognition. *IEEE Trans. Speech and Audio Processing*, 1(6), 1998.

[13] Yanli Zheng and Mark Hasegawa-Johnson. Particle filtering appoach to Bayesian formant tracking. In *IEEE Worksop on Statistical Signal Processing*, 2003.