

COMPARISON OF DISTANCE MEASURES IN DISCRETE SPECTRAL MODELING

Bo Wei and Jerry D. Gibson
 Department of Electrical Engineering,
 Southern Methodist University,
 Dallas, TX 75275

ABSTRACT

We present a general adaptive approach for discrete spectral modeling by minimizing different spectral distances in an adaptive filtering context. By comparing the steady-state error for several spectral distance measures for real speech, we study the performance of these important distance measures. We also present a fast converging algorithm for the COSH distance that is shown to yield a more accurate estimate of the spectral envelope than the Itakura-Saito (I-S) distance measure.

1. INTRODUCTION

Distance measures have been applied to speech processing [1][2] and speech recognition [3] fields and produce measurements of dissimilarity of two speech spectra. Criteria often used for choosing these different distance measures are mathematical tractability and perceptual significance. Some popular distance metrics include the root mean square (RMS) log spectral distance, the Itakura-Saito (I-S) distance, the Likelihood ratio distance, the Cepstral distance, and the COSH distance.

Based on perceptual considerations, it has been shown that good performance can be obtained by using the RMS distance in speaker identification [4]. Gray and Markel suggest that the COSH measure is a better choice where large differences are expected [2].

This paper proposes a new approach to comparing these distances for discrete spectral modeling of speech. Discrete spectral modeling tries to obtain a matching envelope for the speech spectrum. The problem is how to find a set of all-pole model coefficients so that the distance between the model spectrum $\hat{P}(w)$ and original speech spectrum $P(w)$ can be minimized at the harmonic peaks located at w_m , $m=1, \dots, N$, where

$$\hat{P}(w) = \frac{1}{|A(w)|^2} = \frac{1}{\left| \sum_{k=0}^p a_k e^{-jwk} \right|^2} \quad (1)$$

It has been demonstrated that differences between two spectra in formant locations and formant bandwidths cause phonetic differences [3]. That is, a better speech spectrum envelope implies better subjectively meaningful performance.

We compare these distance measures using Least-Mean-Square (LMS) adaptive filtering. By adapting the model coefficients along the error surface of each distance measure, we demonstrate algorithm convergence and evaluate steady-state error. We show that the COSH distance measure gives the best spectral estimate for real speech.

Also, in order to simplify the adaptive procedure for the COSH distance measure, we present a new algorithm that converges much faster than LMS and gives the unique optimum solution. We use the discrete version of all of the distance measures here in discrete spectral modeling.

The paper is organized as follows. In Section 2, we give a brief review of discrete spectral all-pole modeling, and then we propose the LMS approach for comparing the performance of different distance measures in Section 3. The resulting spectral envelopes and the convergence curves are given in Section 4 to show the difference among different distance measures. A new COSH algorithm is introduced in Section 5 and conclusions are given in Section 6.

2. DISCRETE SPECTRAL ALL-POLE MODELING

Frequency-domain based harmonic coding methods, such as Sinusoidal Transform Coding (STC) [5] and Multiband Excitation Coding (MBE) [6], require an accurate model for the spectral envelope of the speech spectrum at the discrete harmonic peaks. A classic spectral estimation method is Linear Prediction (LP). Makhoul [7] proves that the LP error criterion is equivalent to minimizing

$$E_{LP} = \frac{1}{N} \sum_{m=1}^N \frac{P(w_m)}{\hat{P}(w_m)} \quad (2)$$

for the discrete spectrum. Makhoul also notes that the drawback of LP is that spectral estimates are always biased towards the pitch harmonics and this is inherently related to its error criterion. For example, its "error cancellation" and "asymmetry" properties prevent the model spectrum from exactly matching the harmonic peaks at each point.

To overcome these shortcomings, McAulay [8] and El-Jaroudi [9] derive iterative algorithms by minimizing the discrete version of the Itakura-Saito (I-S) distance measure:

$$E_{IS} = \frac{1}{N} \sum_{m=1}^N \frac{P(w_m)}{\hat{P}(w_m)} - \log \frac{P(w_m)}{\hat{P}(w_m)} - 1 \quad (3)$$

An iterative procedure is needed because solving $\partial E / \partial a_i = 0$ leads to a set of nonlinear equations. Discrete All-Pole modeling (DAP) [9], which was introduced in El-Jaroudi's paper, can be shown to result in a better spectral envelope that has better matching conditions than LP. The I-S distance measure is asymmetric, in that, at each point w_m , the contribution to the total error is more significant when $P(w_m)$ is greater than $\hat{P}(w_m)$ than when $P(w_m)$ is smaller [10]. This asymmetry

property performs well when finding the spectral envelope for continuous spectra because it gives the resulting envelope a “sit on top” [8] effect, i.e. it matches spectral peaks better than valleys. For the discrete spectrum case, the RMS distance measure and the COSH are possible alternatives. These distance measures are given by

$$E_{rms} = \frac{1}{N} \sum_{m=1}^N \left[\log \frac{P(w_m)}{\hat{P}(w_m)} \right]^2 \quad (4)$$

$$E_{cosh} = \frac{1}{2N} \sum_{m=1}^N \left[\frac{P(w_m)}{\hat{P}(w_m)} - \log \frac{P(w_m)}{\hat{P}(w_m)} + \frac{\hat{P}(w_m)}{P(w_m)} - \log \frac{\hat{P}(w_m)}{P(w_m)} - 2 \right] \quad (5)$$

From (4), We can see that E_{rms} has the nice properties of symmetry and no error cancellation. The COSH distance also satisfies the symmetry property.

However, minimizing E_{rms} yields a set of nonlinear equations for which it is hard to find a closed form solution. Galas [11] derived a new algorithm for discrete spectral estimation by minimizing the cepstral distance measure

$$E_{cep} = \sum_{n=-\infty}^{\infty} (c_n - \hat{c}_n)^2 \quad (6)$$

where c are the cepstrum coefficients. It can be shown that the truncated cepstral distance is almost identical to the RMS distance measure. Thus, this is an alternative way of using the RMS distance measure. However, the cepstral method requires a relatively large number of coefficients, so we continue to focus on the all-pole model.

We thus find that the different distance measures like LP distance (2), I-S distance (3) and cepstral distance (5) lead to different spectral estimates, which produce different accuracy for their formant estimates. Since differences between two spectra in formant locations and formant bandwidths cause phonetic differences [3], we see that which distance measure is minimized affects perceptual performance.

It is interesting to compare these several distance measures in terms of discrete spectra modeling. However, a major difficulty that prevents us from deriving an algorithm for each distance measure is that minimizing these error criteria does not guarantee a good matching condition or a closed form algorithm. In next section, we use an adaptive filtering method to obtain a general approach that works for all distance measures in discrete spectral estimation.

3. LMS APPROACH FOR DISCRETE SPECTRAL ALL-POLE MODELING

If we think of the distance measures as the “cost function” in terms of adaptive filter terminology, the derivatives of the distance measures will be different gradients. With this interpretation, we use the methods from adaptive filter theory to derive spectral envelopes by minimizing each distance measure. The Least-Mean-Square (LMS) algorithm is easy to implement and the updating of AR coefficients is simple:

$$a_i(n+1) = a_i(n) - \mu \frac{\partial J(n)}{\partial a_i(n)} \quad (7)$$

where a_i $i=0\dots p$ are the AR coefficients. As a function of a ,

$J(n)$ is the cost function in adaptive filtering and depends on the particular distance measure here, and $\partial J(n)/\partial a_i(n)$ is the gradient. Then the spectrum estimation problem using one distance measure becomes the problem of adapting the model coefficients a along the direction of its maximum gradient. One thing we should note here is that the cost function in LMS must have the following characteristics:

1. It reaches its minimum when the system converges to its optimal solution.
2. It must be greater than or equal to zero.

Here, we define the function

$$e(m) = \log \frac{P(w_m)}{\hat{P}(w_m)} = \log P(w_m) - \log \hat{P}(w_m) \quad (8)$$

and $J(n)$ is a function of $e(m)$. The plots for all of the cost

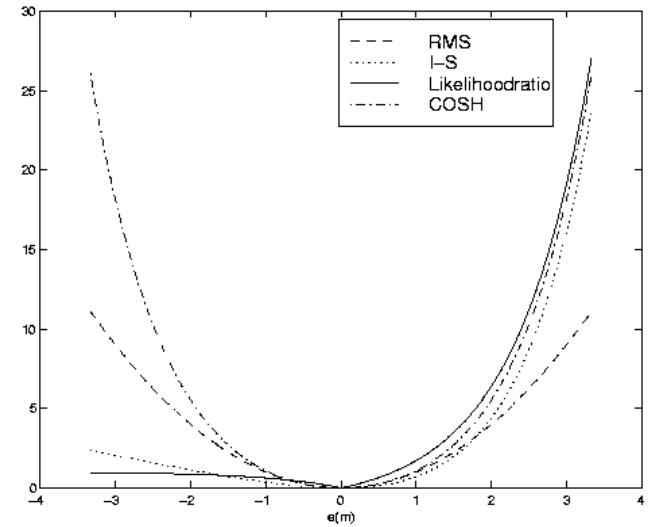


Fig. 1: Cost function for each distance measure

Distance Measures	Gradients $\frac{\partial J}{\partial a_i}$ for all $i, i=1\dots p$
Itakura-Saito	$\frac{2}{N} \sum_{k=0}^p a_k \sum_{m=1}^N [P(w_m) - \hat{P}(w_m)] \cos w_m(i-k)$
RMS	$\frac{4}{N} \sum_{k=0}^p a_k \sum_{m=1}^N \left[\log \frac{P(w_m)}{\hat{P}(w_m)} \hat{P}(w_m) \cos w_m(i-k) \right]$
Likelihood ratio	$\frac{2}{N} \sum_{k=0}^p a_k \sum_{m=1}^N \text{sgn}\left(\frac{P(w_m)}{\hat{P}(w_m)} - 1\right) P(w_m) \cos w_m(i-k)$
COSH	$\frac{2}{N} \sum_{k=0}^p a_k \sum_{m=1}^N \left[\frac{P(w_m)}{\hat{P}(w_m)} - \frac{\hat{P}(w_m)}{P(w_m)} \right] \hat{P}(w_m) \cos w_m(i-k)$

Table 1: Gradient in each LMS algorithm

functions $J(n)$ in term of $e(m)$ are shown in Fig. 1. We can see that all the distance measures we discussed here, I-S, RMS, Likelihood ratio and COSH, satisfy the above properties. We also notice that RMS and COSH cost functions are symmetric while the others are not. The expressions for the gradients in (7) are also listed in Table 1 for each of these distances. We can see that LMS has the same error minimization condition as DAP when we use the I-S distance measure. In another words, DAP is just a special case of the LMS algorithm we propose here by using the I-S distance measure.

4. SIMULATION RESULTS

In the simulations, we perform peak-picking on the spectrum of speech frames(200 samples/frame at 8K samples/s sample rate), obtaining the locations w_m and the magnitudes $P(w_m)$ at the given spectral points. Using the LP coefficients as the initial estimate for a , the adaptation procedures are performed for each distance measure, which leads to the different spectral envelopes in Fig. 2. Also, a convergence comparison is given in Fig. 3 for the same step size μ . Note that RMS distance measure is chosen here as a reference for comparison of estimation error in Fig. 2, i.e.

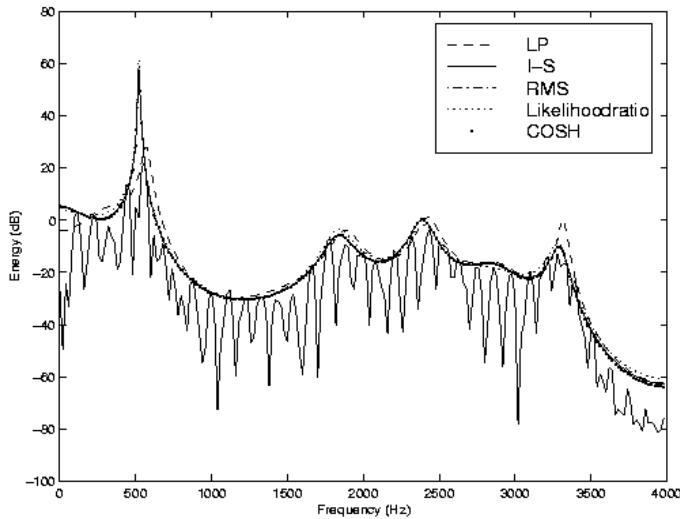


Fig. 2: The resulting spectral envelopes for different distance measures

$$E_{dB} = \sqrt{\frac{1}{N} \sum_{m=1}^N [10 \log_{10} P(w_m) - 10 \log_{10} \hat{P}(w_m)]^2} = \sqrt{E_{rms}} \quad (9)$$

It is clear that in Fig. 3, although I-S, RMS and COSH method have similar steady-state error, COSH converges faster than the others.

From the analysis of steady-state error, COSH gives the minimum error after convergence. Comparing Fig. 3 and Fig. 4,

we see that DAP modeling [9], which has been shown to have optimal convergence using the I-S error criterion, converges to the same point as the Itakura-Saito LMS algorithm. We also derive a quasi-Newton discrete spectral estimation algorithm for the COSH distance measure that converges to the same point as the COSH LMS algorithm, also shown in Fig. 4.

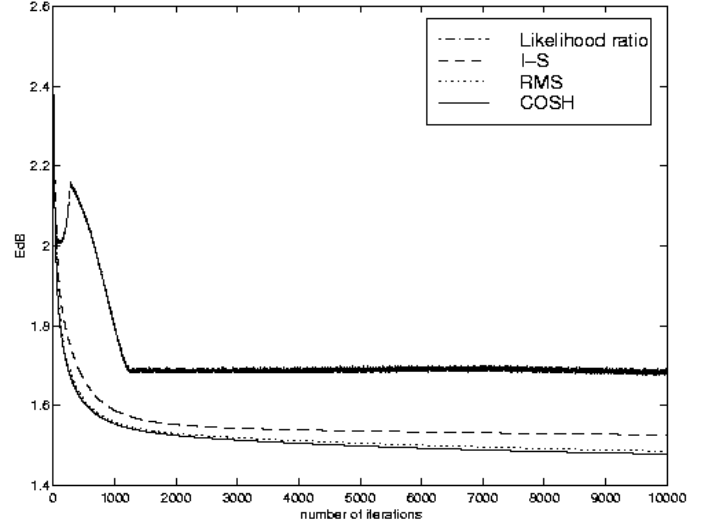


Fig. 3: Convergence comparison for LMS algorithms at the same step size

5. DISCRETE ALL-POLE MODELING USING COSH DISTANCE

By minimizing the COSH distance, we have a similar form to DAP modeling, i.e. similar in terms of minimizing the I-S distance. So we can derive a new discrete all-pole modeling algorithm using COSH distance measure (which we henceforth refer to as the DAPCOSH algorithm). And we can show it will converge to the same steady-state as the LMS algorithm while its adaptation rate is much faster than that of LMS. The convergence comparison result with DAP is given in Fig. 4. We see that the new algorithm converges to the same steady-state as LMS. Taking partial derivatives,

$$\begin{aligned} \frac{\partial E_{cosh}}{\partial a_i} &= \frac{2}{N} \sum_{k=0}^p a_k \sum_{m=1}^N \left[\frac{P(w_m)}{\hat{P}(w_m)} - \frac{\hat{P}(w_m)}{P(w_m)} \right] \hat{P}(w_m) \cos w_m (i-k) \\ &= \frac{2}{N} \sum_{k=0}^p a_k \sum_{m=1}^N \left[P(w_m) - \frac{\hat{P}^2(w_m)}{P(w_m)} \right] \cos w_m (i-k) \quad (10) \end{aligned}$$

we notice that the first term in the equation above is Ra , where

$$R(i) = \frac{1}{N} \sum_{m=1}^N P(w_m) e^{jw_m i} \quad \text{for all } i=1 \dots p \quad (11)$$

and we can construct a P_c from the second item. Let

$$P_c(w_m) = \frac{\hat{P}^2(w_m)}{P(w_m)} \quad (12)$$

so

$$\sum_{k=0}^p a_k \sum_{m=1}^N \frac{\hat{P}^2(w_m)}{P(w_m)} \cos w_m(i-k) = \sum_{k=0}^p a_k \sum_{m=1}^N P_c(w_m) \cos w_m(i-k) = R_c a \quad (13)$$

Now we get a similar formula to DAP,

$$a_{n+1} = R^{-1} R_{cn} a_n = a_n - R^{-1} (R - R_{cn}) a_n \quad (14)$$

We can also prove that R here is a positive-definite matrix, so this algorithm also converges. This algorithm can lead to a better envelope than DAP as can be inferred from the estimation error curves shown in Fig. 4.

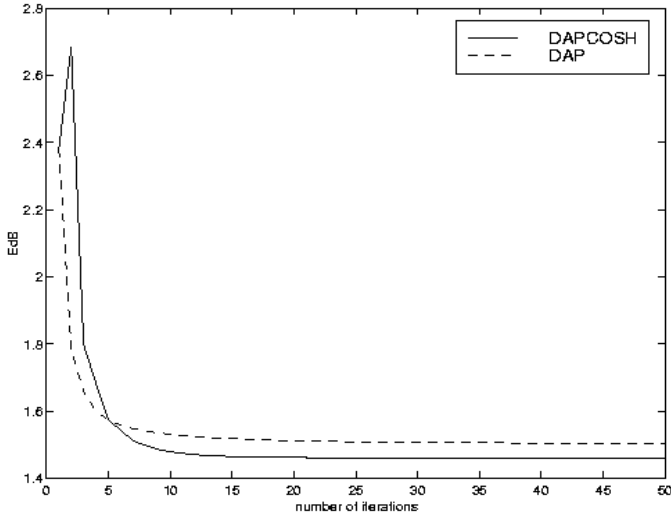


Fig. 4: Convergence of DAP and DAPCOSH

6. CONCLUSION

We have presented here a discussion of several distance measures applied in discrete spectral modeling. For purposes of comparing their performance, we introduce the LMS method from adaptive filter theory. It is shown that the COSH distance measure is the best error criterion among these distances for use in discrete spectral estimation to derive the spectral envelope.

We show it is possible to increase the convergence speed for some of the distance measures. Like DAP modeling, we also propose a new quasi-Newton algorithm DAPCOSH by minimizing the COSH distance measure. The same steady-state error and much faster convergence speed than the LMS algorithm are shown.

REFERENCE

- [1] R. M. Gray, A. H. Gray, Jr., and Y. Masuyama, "Distortion measures for speech processing," *IEEE Trans. Acoustics, Speech, Signal Processing*, ASSP-28 (4) pp. 367-376, August 1980.
- [2] A. H. Gray, Jr. and J. D. Markel "Distance measures for speech processing," *IEEE Trans. Acoustics, Speech, Signal Proc.*, ASSP-28 (4) pp. 380-391, October 1976.
- [3] L. Rabiner and B. H. Juang, *Fundamentals Of Speech Recognition*, Prentice Hall, 1993.
- [4] L. L. Pfeifer, "Inverse filter for speaker identification", Speech Communications Res. Lab., Santa Barbara, CA, Final Rep. RADCTR-74-214, 1974.
- [5] R. McAulay and T. Quatieri, "Sinusoidal Coding" in *Speech Coding and Synthesis*, Elsevier, 1995.
- [6] D. W. Griffin and J. S. Lim, "Multiband Excitation Vocoder" *IEEE Trans. Acoustics, Speech, Signal Processing*, vol. 36, no. 8, August 1988.
- [7] J. Makhoul, "Linear Prediction: A Tutorial Review", *Proc. IEEE*, vol. 63, no. 4, pp. 561-580, April 1975.
- [8] R. J. McAulay, "Maximum likelihood spectral estimation and its application to narrow-band speech coding" *IEEE Trans. Acoustics, Speech, Signal Processing*, vol. ASSP-32, no. 2, pp. 243-251, April 1984.
- [9] A. El-Jaroudi and J. Makhoul, "Discrete All-Pole Modeling", *IEEE Trans. Acoustics, Speech, Signal Processing*, pp. 411-423 February 1991.
- [10] J. Makhoul, "Spectral Linear Prediction: Properties and Applications", *IEEE Trans. Acoustics, Speech, Signal Processing*, vol. ASSP-32, no. 3, pp. 283-296 June 1975.
- [11] T. Galas and X. Rodet, "An improved cepstral method for deconvolution of source-filter systems with discrete spectra: Application to musical sounds" *Proc. Of International Computer Music Conference*, Glasgow, pp. 82-84, 1990.