VOICED/UNVOICED CLASSIFICATION OF SPEECH
WITH APPLICATIONS TO THE U.S. GOVERNMENT LPC-10E ALGORITHM

Joseph P. Campbell, Jr., U.S. Department of Defense
Thomas E. Tremain, U.S. Department of Defense
Department of Defense
Fort Meade, Maryland
U.S.A. 20755-6000

ABSTRACT

This paper describes the development and application of a new voicing algorithm used in the 2400 bit per second U.S. Government’s Enhanced Linear Predictive Coder (LPC-10E). Correct voicing is crucial to perceived quality and naturalness of LPC systems and therefore to user acceptance of LPC systems.

This new voicing algorithm uses a smoothed adaptive linear discriminator to classify the signal as voiced or unvoiced speech. The classifier was determined using Fisher’s method of linear discriminant analysis. The voicing decision smoother is a modified median smoother that uses both the linear discriminant and speech onset to determine its smoothing. The voicing classifier adapts to various acoustic noise levels and features a powerful new set of signal measurements: biased zero crossing rate, energy measures, reflection coefficients, and prediction gains.

The LPC-10E voicing algorithm improves upon other 2400 bps LPC voicing algorithms by providing higher quality synthesized speech. Higher quality is due to halving of the error rate and graceful degradation in the presence of acoustic noise.

INTRODUCTION

Speech production is accomplished by various acoustical excitations of the human vocal tract. The excitation signals arise from vocal fold vibration and vocal tract turbulence. Voiced speech is generated by quasi-periodic vocal fold vibration, only. Unvoiced speech is produced by vocal tract turbulence, only. Mixed (voiced and unvoiced) speech results from simultaneous vocal fold vibration and vocal tract turbulence. Obviously, human speech violates the simple binary voiced/unvoiced voicing hypothesis by being a product of mixed excitation or by changing classes over the classification interval. These violations may be treated as follows.

In an analysis-synthesis system, the goal is to produce the best synthesized speech (highest acceptability or intelligibility). Since the U.S. Government LPC has satisfactory intelligibility, the authors chose to improve LPC acceptability by using a pattern recognition approach for voicing classification based on a training set whose classifications are optimized to synthesize high quality speech. A powerful voicing discrimination parameter set is used to form a linear discriminant classifier.

A major problem for voicing algorithms is the poor performance caused by speech corrupted by acoustic noise. This algorithm was designed to alleviate the problem by adapting the classifier’s coefficients according to the acoustic noise level.

PARAMETER STUDY

The ideal voicing parameter set is independent and discriminates consistently between voiced and unvoiced classes for a wide range of speakers in various acoustic noise environments. Over 100 measurements and transforms were investigated. Investigation of the transforms sometimes revealed the linear separation necessary for linear discrimination. Many of the voicing parameters selected were used in the previous U.S. Government LPC voicing classifier [1] or were suggested by others [2,3,4,5,6,7]. However, the application of IVRC2, R0, and aR1 is novel to voicing classifiers.

Parameter selection was based on the authors’ knowledge of human speech production, class separation (discussed later), and detailed graphic analysis. The graphics tools used included multicolor parameter tracks, discriminated histograms and scatter plots. Figures 1 and 2 display the analysis systems used by the authors: LPC GRAPH, which is integrated in the LPC-10E development system, and SPEASY2 (Speak Easy 2).

FISHER’S METHOD [8]

As demonstrated by other voicing algorithms, linear discrimination is appropriate for voicing classification [1, 6]. For the two-class voiced/unvoiced problem, using N parameters, the M-level adaptive linear discriminant classifier is:

\[ \sum_{i=1}^{N} a_{ij} p_i + c_j > 0 \]

and where \( a_{ij} \) and \( c_j \) are weights and \( p_i \) is the measured signal parameter. The discriminator classifies the speech segment as voiced if the expression is true; otherwise, classify it as unvoiced. The \( a_{ij} \)'s and \( c_j \)'s adapt to an acoustic noise level by selecting \( j \) according to the signal-to-noise ratio (SNR).

One technique to determine the weights \( a_{ij} \) and \( c_j \) is Fisher’s Optimal Linear Discriminant Analysis method. It is optimal under multivariate-normal and equal covariance parameter sets. Fisher’s method chooses the coefficients that maximize the ratio of the difference of the means of the linear combination in the two groups to their common variance. Although this method assumes equal covariance, in our case, the matrices are close enough that it makes little or no difference in the results to assume equality. In addition, this method is quite robust to non-normality.

Fisher’s method requires a set of training data consisting of properly classified groups (hand-painted voicing decisions) and their corresponding parameters. The computer program used by the authors determines the weights \( a_{ij} \) and \( c_j \) and the class separation Mahalanobis’ D2 (the difference between the group means of the discriminant function corrected for correlation effects).

The Mahalanobis D2 provides a very useful quantitative measure of classifier performance for the given parameter set. This aids in parameter and transform selection and in refinement of the calculation of parameters.

U.S. GOVERNMENT LPC-10E VOICING

The LPC-10E voicing algorithm makes voicing decisions for each 11.25-ms half frame of input speech using the following signal measurements: zero crossing rate, energy measures, reflection coefficients (RC's), and prediction gains. The classifier adapts to varying acoustic noise levels and was determined using Fisher’s method of linear discriminant analysis.
Figure 1. LPC GRAPH's display of input speech ("Mabel stood on the rock"), and LPC-10E's voicing window placement, preliminary and smoothed voicing, speech onsets, voicing discriminant classifier, and voicing parameters.

Figure 2. SPEASY2's discriminated histograms of LPC-10E voicing parameters and unsmoothed classifier for 10 sentences. The empty and solid bins represent voiced and unvoiced speech, respectively.

9.11. 2
The voicing algorithm consists of the following operations performed in sequence:

1) Input speech filtering
2) Voicing parameter measurement
3) Voicing classification
4) Smoothing of voicing classifications

Input Speech Filtering

The 8-kHz sampled digital input signal is low-pass filtered to extract the predominantly low-band voiced excitation signal. The low-pass filter's characteristics were selected to provide maximum discriminator function class separation. The filter is a 19-point FIR equiripple linear-phase design having an 800-Hz passband with 1-dB ripple.

Voicing Parameter Measurement

Except as noted, the following signal parameters are calculated for each half frame voice decision for use by the voicing classifier. Zero crossings and energy correlation measures are calculated on low-band and full-band speech. The voicing window is located only where the speech waveform is stationary or, equivalently, to avoid placement of the window where the speech waveform is rapidly changing in character. Each measure is taken over either the first or second half of the voicing window; this determines the length of the summations in the expressions below to be between 45 and 78 samples. In the following equations, z represents the 100-Hz high-pass filtered input signal and x represents the low-band 800-Hz low-pass and 100-Hz high-pass filtered input signal. (Refer to figures 1 and 2 for the behaviors of the following voicing parameters.)

Low-Band Speech Energy. Energy is the most obvious and simple indicator of "voicedness." Typically, voiced sounds have 30-dB greater low band energy than unvoiced sounds. As shown below, E is the normalized low-band energy. The normalization makes E reliable for different speakers and different acoustic environments. Energy estimation by sum of magnitudes exhibits greater class separation under Fisher's method than by root mean square (rms) or logarithm of rms. A speaker-dependent and stable-energy reference is the average low-band voiced energy <LBVE> implemented by a moving-average, first-order recursive filter whose input is the low-band energy during voiced half frames.

\[ E = \frac{\sum |z_i|}{<LBVE>} \]

Zero Crossing Rate. The zero crossing rate (ZC) counts sign changes in the input speech signal s. This indicates the dominant spectral concentration in the speech signal. A high ZC indicates that high frequencies dominate s, while a low ZC indicates low frequency dominance. Voiced speech's quasi-periodic excitation signal produces a concentration of low frequency energy, resulting in a low zero crossing rate. Unvoiced speech tends to have a high zero crossing rate due to the preemphasized vocal tract response.

Dither is added to the signal before the zero crossings are counted to guard against low-level input and low-band corruption of the input speech (i.e., ac main's interference and fan noise). The dither signal is a half-sample-rate square wave whose amplitude adapts to the low-band voiced and unvoiced energies.

First Reflection Coefficient. The first reflection coefficient, RC1, is calculated explicitly as the normalized short-term autocovariance coefficient at unit sample delay. This parameter measures spectral tilt over the entire speech passband. Voiced excitation produces speech that is highly correlated between adjacent samples with significant spectral tilt of decreasing magnitude with increasing frequency. Unvoiced speech typically lacks this quality.

\[ RC_1 = \frac{\sum z_i z_{i-1}}{\sum z_i^2} \]

Preemphasized Energy Ratio [4]. QS is the ratio of the energies in the 6-dB preemphasized first-order difference signal to the regular signal. This parameter measures weighted high-band energy and is similar to the first RC; however, it is less sensitive to spectral tilt at the band edges. It is well behaved in detecting high energy unvoiced plosives, which cause most voicing classifiers to declare voiced.

\[ Q_S = \sum \frac{|s_i - s_{i-1}|}{\sum |s_i|} \]

Second Reflection Coefficient [5]. The second reflection coefficient is a measure of the relative degree of spectral peak or Q. IV/RC2 is restricted to indicate low-band Q by being computed on 800-Hz low-pass filtered, 4:1 decimated speech (§). IV/RC2 is usually negative and increases in magnitude with increasing Q. Voiced speech tends to have a very significant low-band spectral peak and therefore an IV/RC2 close to -1. Unvoiced speech generally lacks a low-band spectral peak, yielding an IV/RC2 near zero. This parameter is calculated over a fixed window every 22.5 ms for use by the pitch tracker's inverse filter.

\[ IV/RC_2 = r_2 - r \]

Causal Pitch Prediction Gains. The product of the first-order forward and reverse causal pitch prediction gains is shown below. This crosscorrelation at minus a pitch epoch lag is a strong indicator of backward-looking pitch periodicity. Pitch measures generally respond slowly because they require relatively long analysis windows. Since steady-state voiced or unvoiced intervals are easily classified by the above parameter set, pitch measures are usually of little use. However, the authors have found aR1 very useful in detecting pitch trailing off in the speech waveform, although it is not particularly stable. This results in a major improvement over many other classifiers: the ability to properly classify trailing off voiced to unvoiced transitions. In the following expressions, r is the pitch period determined by the absolute magnitude difference function (AMDF).

\[ a R_0 = \left( \frac{\sum z_i z_{i+r}}{\sum z_i^2} \right)^2 \]

Noncausal Pitch Prediction Gains. The product of the first-order forward and reverse noncausal pitch prediction gains is shown below. This crosscorrelation at plus a pitch epoch lag is a strong indicator of forward-looking pitch periodicity. The authors have found aR1 very useful in detecting the onset of pitch and therefore voicing in the speech waveform, although it is not extremely stable. Both aR0 and aR1 approach unity for voiced speech due to adjacent pitch epoch similarity and approach zero for unvoiced speech.

\[ a R_f = \left( \frac{\sum z_i z_{i+r}}{\sum z_i^2} \right)^2 \]

Voicing Classification

Adapting the Classifier to Acoustic Noise. In order to select the appropriate voicing decision weights, aR0 and aR1, the first step of the voicing classifier is to determine f by estimating the SNR as the ratio of the average full-band voiced energy to the average full-band unvoiced energy. The SNR estimate is robust because it uses reliable, accurate energy measures. To avoid contamination of the unvoiced energy estimate from high-energy voicing misclassifications, the input to the unvoiced energy averaging filter is limited to less than 10 dB of the previous input.

Voicing Discriminant. The voicing parameters from the above calculations are used in the LPC-10E voicing discriminator. Tentatively
classify the half frame as voiced if the following expression holds true, otherwise declare it unvoiced:

\[ \sum_{i=1}^{N} a_{i,j} p_i + c_j > 0 \quad \text{where} \quad j \in \{0, \ldots, M-1\} \]

and where \( N = 7 \), \( M = 8 \), and \( a_{i,j} \) and \( c_j \) are specified in the U.S. Government LPC-10E FORTRAN Program. For example, in non-noisy input speech, let \( j = 0 \); the LPC-10E voicing classifier for this case is:

\[
p_1 = [E, ZC, RC_1, Q_5, JV, RC_2, \alpha R_0, \alpha R_f] \\
\alpha_{1,0} = [1158, -108, 832, -4096, -1018, 1195, 1011] \\
c_0 = 3462
\]

Smoothing of Voicing Classifications

Tentative voicing decisions are made two frames in the future for each half frame. These decisions are carried through one frame in the future to the present frame, where they are examined and smoothed, resulting in the final voicing decisions for each half frame.

Voicing classification irregularities are smoothed by the modified median smoother. To determine how strongly voiced or unvoiced a signal is, the smoother uses the voicing discriminant function. The smoothing is further modified if a speech onset (a small portion of the speech where its character is rapidly changing) and a voicing decision transition occur within one half frame. In this case, the voicing decision transition is extended to the point of speech onset. For transmission purposes, there are constraints on the minimum duration and transition of voicing decisions. The smoother takes these constraints into account.

RESULTS

This new LPC-10E voicing algorithm is superior to other 2400 bps LPC-based voicing algorithms because it provides improved synthesized speech quality in both noisy and non-noisy environments. It has a more graceful degradation in the presence of noise and, depending on the speech used, allows half the error rate in noiseless speech, compared to the previous algorithm. These improvements are the result of the application of Fisher's method and sophisticated graphics tools to the voicing problem and the discovery of a powerful new voicing parameter set. Separately, the authors' voicing parameters are not completely reliable; however, their combination is synergistic. The new voicing algorithm correctly classifies typically difficult sounds, such as high energy unvoiced plosives and trailing off voiced to unvoiced transitions.

Many LPC voicing algorithms are intentionally biased to favor voiced classifications (due to unequal misclassification cost). This results in the characteristic LPC "buzziness" due to incorrect application of voiced periodic excitation to unvoiced sounds in the LPC synthesizer. The new LPC-10E voicing algorithm classifies unvoiced speech more accurately and is unbiased, producing speech with greater clarity and quality. For one case of 27 seconds of speech (2400 voicing decisions), the misclassification rates are 3 and 5 percent, respectively, for the new and previous algorithms. Most importantly, the new voicing algorithm results in higher quality synthesized speech and improved user acceptability of LPC-10E based systems.

ACKNOWLEDGMENTS

The authors’ high resolution color graphics tools were provided by Dave Andre and Craig Reese.

We are grateful to DCA, Vishu Viswanathan and Bill Russell of BBN, and George Doddington and Joe Ficione of TI for the use of their training data bases.

We wish to express our gratitude to the following individuals for their useful and stimulating suggestions: Dave Andre, Bishnu Atal, Ron Cohn, Rich Dean, Bruce Fette, George Kang, Dave Kemp, J. David Murley, Doug Rahikka, Leah Siegel (Jamieson), Steve Sroka, and Shawn Campbell.

REFERENCES