[5] M. Benidir and B. Picinbono, "Extensions of the stability criterion for ARMA filters," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-35, pp. 425-431, Apr. 1987.

[6] Y. Bistritz, "A circular stability test for general polynomials," *Syst. Contr. Lett.*, vol. 7, pp. 89-97, 1986.

[7] M. Benidir and B. Picinbono, "Comparison between some stability criteria of discrete-time filters," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 36, pp. 993-1001, July 1988.

[8] H. Krishna, B. Krishna, and S. D. Morgera, "Efficient procedure for stepup and stepdown computations," in *Proc. ICASSP'88*, New York, NY, Apr. 1988, pp. 1651-1654.

Silent and Voiced/Unvoiced/Mixed Excitation (Four-Way) Classification of Speech

D. G. CHILDERS, M. HAHN, AND J. N. LARAR

Abstract—**We present an algorithm for automatically classifying speech into four categories: silent and speech produced by three types of excitation, namely, voiced, unvoiced, and mixed (a combination of voiced and unvoiced). The algorithm uses two-channel (speech and electroglottogram) signal analysis and has been tested on data from six speakers (three male and three female), each speaking five sentences. An overall correct classification accuracy of approximately 98.2 percent was achieved when compared to skilled manual classification. This is superior to previously reported automatic classification schemes. If word boundary errors, including the beginning and ending of sentences, are excluded, then the algorithm's performance improves to 99.5 percent.**

Introduction

In previous work [1], we described a two-channel, two-way (V/U-S) algorithm for automatically classifying speech. This algorithm used the speech and electroglottogram (EGG) signals. One of our objectives has been to demonstrate that two-channel-based algorithms can lead to computational and performance improvements over algorithms based on acoustic-signal-only analysis methods. We recognize that in many situations, the EGG signal is either unavailable or cannot be used. However, both the speech and EGG signals can be used in the laboratory to help benchmark the performance of numerous speech systems. We advocate this approach.

Previous research has focused on three-way speech classification, i.e., either V/U/S or V/U/M [2]-[13]. The speech classification problem is important because its solution affects other speech analysis, synthesis, and recognition problems. For example, speech classification can help reduce the number of lexical candidates in speech (word) recognition, improve speech synthesis by selecting the proper excitation, and improve the performance of phoneme boundary detection in speech analysis. Consider the large vocabulary isolated word recognition problem. By using only four-way (V/U/M/S) classification and stress analysis, one can define an equivalence class of words having the same representation or "coding" [9], [10], [14]. For example, the words speed, steep, scout, and stop all belong to the same equivalence class of U/S/stressed-U/V/U. In an isolated word recognition system, the search to identify a test word among all possible candidates can be reduced by using such a simple coding technique. Following such a reduction of the lexical candidates, one may perform other, more detailed analyses to match the test word with one of the remaining words.

Some of the problems with classifying speech as V/U using acoustic-signal-based algorithms are caused by the use of a large analysis frame, a low level of voicing, or even the strength of the first formant. Classification of U/S segments is even more difficult for such algorithms. Typically, researchers have adopted sophisticated approaches to overcome these problems, using additional features, a statistical approach, or an optimized set of parameters [2], [3], [7], [12], [13].

A Speech-EGG-Based Algorithm for V/U/M/S Classification

A. Algorithm Overview

The properties and some applications of the EGG to speech analysis appear in [1], [8], [15]-[18]. The EGG offers advantages not readily available from a microphone, even a throat contact microphone. The EGG is not susceptible to environment noise, providing instead a direct measure of vocal fold contact [19], while the throat contact microphone provides an acoustic signal similar in form to that provided by other microphones.

The EGG amplitude varies both within and across speakers. Baseline variations in the EGG may be removed by differentiating the EGG. For voiced segments, the EGG usually has only two zero crossings per fundamental (pitch) period of voicing. One exception is vocal fry. For unvoiced segments, the electroglottograph output is a very low-level high-frequency noise-like signal generated by the internal electronics of the device that is easily distinguished from the excitation for voiced speech. Thus, V/U-S classification is achieved using a combination of EGG amplitude and level-crossing rate [1], [8]. Mixed excitation detection is accomplished by noting that the EGG signal appears similar to that for voiced sounds, but the speech signal is small in amplitude and has a high level-crossing rate (see Fig. 2 and other examples in [1]). Silent intervals are detected by observing that the EGG waveform appears as it does for unvoiced speech and that the speech signal is below a predetermined energy threshold. The two-channel, four-way speech classification algorithm appears in Fig. 1. Note that the algorithm does not use endpoint classification, but this could be added if desired [4], [11].

B. Algorithm Details

Fig. 2 depicts both illustrative results and some difficulties encountered in attempting to evaluate the level-crossing rate (LCR) and energy of the EGG signal. Some fluctuations in the EGG data may be removed by simple differentiation, a procedure which also yields a waveform with enhanced positive and negative peaks. These peaks occur approximately at the instants of glottal opening and closure, respectively [1], [8], [18]. The differentiation is implemented as a backward difference equation. The differentiated EGG is then normalized by dividing by its maximum positive value. The resulting waveform is denoted as the DEGG and is shown at the bottom of Fig. 2.

The two-way, V-M and U-S classification uses the LCR and energy information from the DEGG as follows. The DEGG is segmented into 10 ms frames of 100 samples each (10 kHz sampling

Manuscript received March 8, 1988; revised February 13, 1989. This work was supported in part by NSF Grant ECE-8413583, NIH Grants NINCDS RO1 NS17078 and NS 27022, the University of Florida Center of Excellence Program in Information Transfer and Processing, the Florida High Technology and Industry Council under Grant 4505208-12, and Mind-Machine Interaction Research Center.

D. G. Childers and M. Hahn are with the Department of Electrical Engineering, University of Florida, Gainesville, FL 32611.

J. N. Larar was with AT&T Bell Laboratories, Murray Hill, NJ. He is now with the Department of Medicine, University of Miami, Miami, FL 33101.

IEEE Log Number 8930546.

0096-3518/89/1100-1771$01.00 © 1989 IEEE

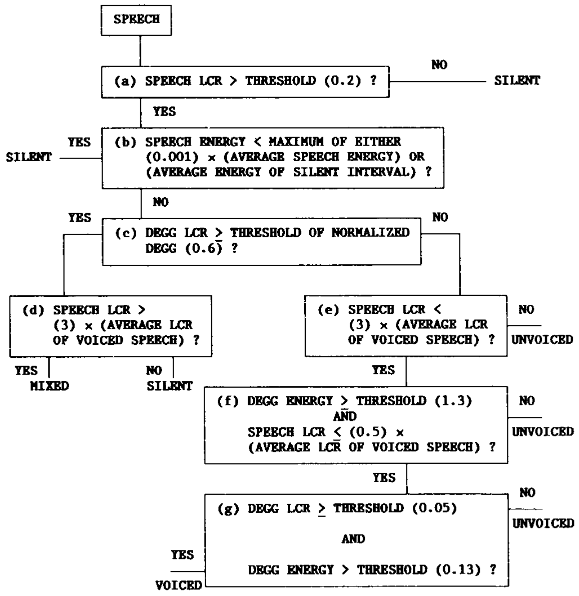
SPEECH
(a) SPEECH LCR > THRESHOLD (0.2) ?
NO
SILENT
YES
(b) SPEECH ENERGY < MAXIMUM OF EITHER (0.001) × (AVERAGE SPEECH ENERGY) OR (AVERAGE ENERGY OF SILENT INTERVAL) ?
YES
SILENT
NO
(c) DEGG LCR ≥ THRESHOLD OF NORMALIZED DEGG (0.6) ?
YES
NO
(d) SPEECH LCR > (3) × (AVERAGE LCR OF VOICED SPEECH) ?
YES
MIXED
NO
SILENT
(e) SPEECH LCR < (3) × (AVERAGE LCR OF VOICED SPEECH) ?
NO
UNVOICED
YES
(f) DEGG ENERGY ≥ THRESHOLD (1.3) AND SPEECH LCR ≤ (0.5) × (AVERAGE LCR OF VOICED SPEECH) ?
NO
UNVOICED
YES
(g) DEGG LCR ≥ THRESHOLD (0.05) AND DEGG ENERGY > THRESHOLD (0.13) ?
NO
UNVOICED
YES
VOICED

Fig. 1. Speech-EGG algorithm for V/U/M/S (four-way) classification of speech.

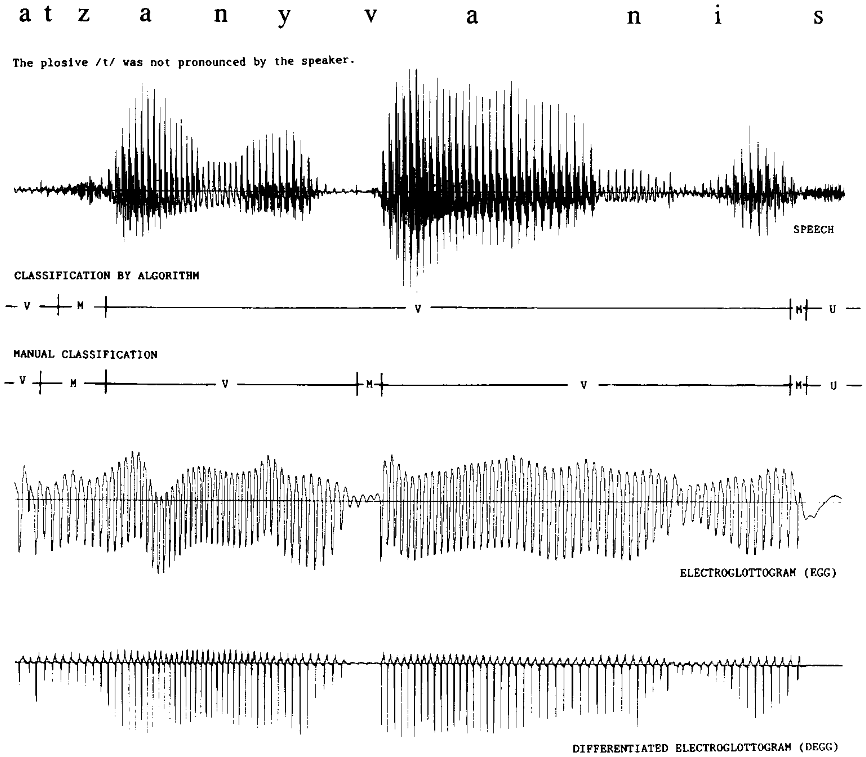

Fig. 2. Comparison of V/U/M/S classification by the algorithm and manual procedures. The speech is a segment of the sentence "That zany van is azure" spoken by a male subject.

rate). The energy of each frame n of the DEGG signal is given by

$$E_D(n) = \sum_{i=1}^{100} \left(D((n-1)\,100 + i)\right)^2 \qquad (1)$$

where $D(j)$ denotes the sample value of the DEGG signal. The LCR of the DEGG is calculated at the -0.5 level, also on a frame basis. The energy and LCR contours are smoothed with a three-point zero-phase filter with coefficients (0.19, 0.62, 0.19), such that the filter output is $y_n = 0.19x_{n-1} + 0.62x_n + 0.19x_{n+1}$. The energy and LCR for the speech signal are calculated in a similar manner and the contours are also smoothed.

Other calculations include the following:

1) average energy levels for both the speech and DEGG signals for voiced segments,

2) average of the rectified voiced speech,

3) average LCR for voiced segments that exceed the 10 percent level of signal calculated in step 2) above.

4) smooth the preliminary four-way classification to remove obvious errors in a string. For example, if the preliminary classification is . . . VVVUVV . . ., then "smooth" this string to give . . . VVVVVV . . . .

Various threshold values are determined empirically, but are fixed once selected. The thresholds we selected are shown in Fig. 1 enclosed in the parentheses. Samples of silent intervals are required to establish several threshold values. The other numbers in Fig. 1 were determined experimentally to establish the decision levels for steps (b), (d), and (f).

Results and Discussion

The algorithm has been tested with data from six speakers (three male and three female), each speaking five sentences. The sentences were as follows.

1) We were away a year ago. (Voiced.)

2) Early one morning a man and a woman ambled along a one mile lane. (Voiced and nasals.)

3) Should we chase those cowboys? (Fricatives and plosives.)

4) That zany van is azure. (Voiced fricatives, i.e., mixed.)

5) We saw the ten pink fish. (Unvoiced plosives and fricatives.)

This data set is more extensive than that used in [2], [3], [7], [13].

The threshold values were determined with data from two speakers (one male and one female), each speaking the first three sentences only. An example comparing the algorithm classification to manual classification is shown in Fig. 2. Note that the classification task is aided by the EGG and the DEGG.

The detailed classification results appear in Table I. The designation for the test data sets is as follows.

Complete: Refers to all the data from all six speakers.

Threshold: The subset of Complete used to establish the threshold values, one male and one female speaker, for the first three sentences only.

Nonthreshold: The subset of Complete not included in the Threshold set.

Male: The male speaker subset of Complete.

Female: The female speaker subset of Complete.

The overall correct recognition rate is 98.2 percent when compared to manual classification of the data. The recognition rate is an improvement over the overall 95 percent rate reported in [3], [7] and 88 percent reported in [2], [13]. Nearly 83 percent correct classification of the mixed excitation frames was achieved in [7], which we have increased to 89 percent.

Table II provides a breakdown of the types of errors. The most troublesome classifications for the algorithm were unvoiced and mixed excitation frames. A large number of errors (38.7 percent) occurred at the beginning and ending of the sentences. If these errors are ignored, then the overall performance of the algorithm becomes 99.23 percent. If we further ignore the errors that occurred at the boundaries between words, then the overall performance increases to 99.5 percent. The major cause of errors at word boundaries (approximately 60 percent), including the beginning and ending of the sentences, was due to U to V and S to V misclassifications. These errors were caused by a failure to properly recognize voice-onset and voice-offset intervals. An algorithm for recognizing voice onset and offset using the EGG is described in [11] and is an extension of the one in [3], [4]. Note that there is a slight tendency for the algorithm to perform better using male speech than female speech.

TABLE I
Classification Results

TEST DATA SETS	TOTAL NUMBER OF FRAMES	NUMBER OF FRAMES IN ERROR	ERROR RATE (%)	CORRECT RATE (%)
COMPLETE	7785	146	1.88	98.12
THRESHOLD	1622	20	1.23	98.77
NON-THRESHHOLD	6163	126	2.04	97.96
MALE	3887	47	1.21	98.79
FEMALE	3898	99	2.54	97.46

TABLE II
Error Analyses in Number of Frames

CLASSIFICATION OUTPUT MANUAL CLASSIFICATION	V	U	M	S	CORRECT RATE (%)
V	5298	24	39	6	98.71
U	20	710	5	6	95.82
M	5	5	81	0	89.01
S	27	9	0	1550	97.73

Conclusions

We advocate the laboratory use of this algorithm to benchmark speech system performance. The benchmarking can be done automatically and the results compared to acoustic-signal-only-based algorithms. A useful improvement to the algorithm would be a diagnostic capability. For example, perhaps the algorithm could identify and label the frames that were particularly difficult to classify. Such information could conceivably be used to improve system designs. We believe a spectral distance metric can be used to improve the V/U (U/V) and V/S (S/V) classifications.

References

[1] A. K. Krishnamurthy and D. G. Childers, "Two-channel speech analysis," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-34, pp. 730-743, Aug. 1986.

[2] B. S. Atal and L. R. Rabiner, "A pattern recognition approach to voiced-unvoiced-silence classification with applications to speech recognition," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-24, pp. 201-212, June 1976.

[3] L. R. Rabiner and M. Sambur, "Application of an LPC distance measure to the voiced-unvoiced-silence detection problems," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-25, pp. 338-343, Aug. 1977.

[4] L. R. Rabiner and R. W. Schafer, *Digital Processing of Speech Signals.* Englewood Cliffs, NJ: Prentice-Hall, 1978.

[5] J. D. Markel and A. H. Gray, Jr., *Linear Prediction of Speech.* New York: Springer-Verlag, 1976.

[6] R. W. Schafer and J. D. Markel, Eds., *Speech Analysis.* New York: IEEE Press, 1979.

[7] L. J. Siegel and A. C. Bessey, "Voiced/unvoiced/mixed excitation classification of speech," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-30, pp. 451-460, June 1982.

[8] D. G. Childers and J. N. Larar, "Electroglottography for laryngeal function assessment and speech analysis," *IEEE Trans. Biomed. Eng.*, vol. BME-31, pp. 807-817, Dec. 1984.

[9] J. N. Larar, "Towards speaker independent isolated word recognition for large lexicons: A two channel, two-pass approach," Ph.D. dissertation, Univ. Florida, Gainesville, 1985.

[10] ——, "Lexical access using broad acoustic-phonetic classifications," *Comput. Speech Language*, vol. 1, pp. 47-59, 1986.

[11] N. B. Pinto, D. G. Childers, and A. L. Lalwani, "Formant speech synthesis: Improving production quality," *IEEE Trans. Acoust., Speech, Signal Processing*, to appear, Dec. 1989.

[12] F. Daaboul and J. P. Adoul, "Parametric segmentation of speech into voiced-unvoiced-silence intervals," in *Proc. IEEE Conf. Acoust., Speech, Signal Processing*, Hartford, CT, May 1977, pp. 327-331.

[13] L. R. Rabiner, C. E. Schmidt, and B. S. Atal, "Evaluation of a statistical approach to voiced-unvoiced-silence analysis for telephone quality speech," *Bell Syst. Tech. J.*, vol. 56, pp. 455-482, Mar. 1977.

[14] D. W. Shipman and V. W. Zue, "Properties of large lexicons: Implications for advanced word recognition systems,"in *Proc. IEEE Conf. Acoust., Speech, Signal Processing*, Paris, France, May 1982, pp. 546-549.

[15] W. Hess, *Pitch Determination of Speech Signals*. New York: Springer-Verlag, 1983.

[16] W. Hess and H. Indefrey, "Accurate pitch determination of speech signals by means of a laryngograph," in *Proc. Int. Conf. Acoust., Speech, Signal Processing*, vol. 1, 1984, pp. 1813.1.1-1813.1.4.

[17] ——, "Accurate time domain pitch determination of speech signals by means of a laryngograph," *Speech Commun.*, vol. 6, pp. 55-58, Mar. 1987.

[18] D. G. Childers and A. K. Krishnamurthy, "A critical review of electroglottography," *CRC Crit. Rev. Bioeng.*, vol. 12, no. 2, pp. 131-164, 1985.

[19] D. G. Childers, D. M. Hicks, G. P. Moore, and Y. A. Alsaka, "A model for vocal fold vibratory motion, contact area, and the electroglottogram," *J. Acoust. Soc. Amer.*, vol. 80, pp. 1309-1320, Nov. 1986.

A Tight Upper Bound of the Average Absolute Error in a Constant Step-Size Sign Algorithm

EWEDA EWEDA

Abstract—**A direct performance index of the adaptive filtering sign algorithm (SA) is the average absolute error (AAE) at the output of the filter. Adopting this performance index, the correspondence achieves an easy analysis of SA under a weak assumption. It is proved, for both deterministic and random inputs of the filter, that the AAE has a tight upper bound that exceeds the minimum AAE by half the product of the step size and power of the filter input. The assumption used is existence of average squared and average absolute values of input signals of the filter. A practical interest of the correspondence is that it provides a formula for the biggest step size as a function of tolerable adaptation noise-to-desired-signal ratio.**

I. Introduction

The correspondence is concerned with analysis of the adaptive filtering sign algorithm (SA) [1]-[5]

$$\boldsymbol{H}_{k+1} = \boldsymbol{H}_k + \mu \operatorname{sgn}(e_k)\boldsymbol{X}_k \tag{1}$$

$$e_k \triangleq a_k - \boldsymbol{H}_k^T\boldsymbol{X}_k. \tag{2}$$

In (1) and (2), $\boldsymbol{H}_k$ is an N-dimensional vector composed of the weights of the adaptive filter at time k, a_k is the desired response of the filter, $\boldsymbol{X}_k$ is an N-dimensional vectorial observation on the basis of which a_k is estimated, e_k is the estimation error, μ is a positive number usually called the step size of the algorithm, and $\boldsymbol{H}_k^T$ is the transpose of $\boldsymbol{H}_k$. Technical simplicity of this algorithm makes it appealing in adaptive filtering applications. The SA may be considered [1], [2] as a stochastic gradient algorithm that searches recursively the vector $\boldsymbol{H}$ that minimizes the mean absolute error $E(|a_k - \boldsymbol{H}^T\boldsymbol{X}_k|)$ where E denotes the mathematical expectation. In applications, the algorithm (1) works with one realization of the random process $(a_k, \boldsymbol{X}_k)$ and it actually searches the vector $\boldsymbol{H}$ that minimizes the average absolute error (AAE) defined by

$$\epsilon(\boldsymbol{H}) \triangleq \lim_{k\to\infty} \frac{1}{k}\sum_{j=1}^{k} |a_j - \boldsymbol{H}^T\boldsymbol{X}_j|. \tag{3}$$

Therefore, a direct performance index of this algorithm for a given value of μ is the actual AAE, ϵ_μ, at the output of the adaptive filter defined by

$$\epsilon_\mu \triangleq \lim_{k\to\infty} \sup \left(\frac{1}{k}\sum_{j=1}^{k} |e_j|\right). \tag{4}$$

As will be seen in the correspondence, adopting this performance index enables easy analysis of the SA under a weak assumption. This is not the case with analyses adopting the mean-squared performance index [3], [4] where restrictive assumptions are used such as mutual independence of successive vectors $\boldsymbol{X}_k$.

In this correspondence, we derive an upper bound of ϵ_μ. It is proved in Section II that

$$\epsilon_\mu \le \epsilon_{\min} + \tfrac{1}{2}\mu P \tag{5}$$

where

$$\epsilon_{\min} \triangleq \min_{\boldsymbol{H}} (\epsilon(\boldsymbol{H})) \tag{6}$$

is the minimum average absolute error and

$$P \triangleq \lim_{k\to\infty} \frac{1}{k}\sum_{j=1}^{k} \|\boldsymbol{X}_j\|^2 \tag{7}$$

is the average power of the filter input. The norm $\|\boldsymbol{X}\|$ of a vector $\boldsymbol{X}$ is defined as $\sqrt{\boldsymbol{X}^T\boldsymbol{X}}$. The only assumption used to prove (5) is the existence of the limits in (3) and (7) which is a weak assumption. The inequality (5) holds for all values of μ. It is also valid for both deterministic and random inputs of the filter. In the random case, existence of the limits (3) and (7) is a weak ergodicity assumption that admits strong correlation in input signals of the filter (Section II). In Section III, it is shown that the bound (5) is tight for all values of μ. Simulation results, given in Section IV, agree with theoretical results of Sections II and III. A useful conclusion that can be obtained from (5) is that the excess average absolute error $\epsilon_\mu - \epsilon_{\min}$, which is a measure of the adaptation noise, is proportional to μ. A practical interest of the correspondence is that it gives a formula, (21) below, for the biggest step size sufficient to keep the adaptation noise-to-desired-signal ratio below a tolerable level.

II. Derivation of the Bound

It is easy to show that the function $\epsilon(\boldsymbol{H})$ is convex-up in $\boldsymbol{H}$ [1], i.e.,

$$\epsilon(\alpha\boldsymbol{H}_1 + (1-\alpha)\boldsymbol{H}_2) \le \alpha\epsilon(\boldsymbol{H}_1) + (1-\alpha)\epsilon(\boldsymbol{H}_2), \quad 0 \le \alpha \le 1.$$

Consequently, $\epsilon(\boldsymbol{H})$ has at least one minimum. Due to convexity, all minima yield the same value, $\epsilon_{\min}$, of $\epsilon(\boldsymbol{H})$. Thus, if $\boldsymbol{H}_*$ is a minimum of $\epsilon(\boldsymbol{H})$, then

$$\epsilon(\boldsymbol{H}_*) = \epsilon_{\min}. \tag{8}$$

Manuscript received May 3, 1988; revised January 24, 1989.

The author is with the Military Technical College, Kobry El-Koppa, Cairo, Egypt.

IEEE Log Number 8930547.

0096-3518/89/1100-1774$01.00 © 1989 IEEE