VOICING STATE DETERMINATION OF CO-CHANNEL SPEECH

Daniel S. Benincasa
Rome Laboratory/OCSS
26 Electronic Pkwy
Rome, NY 13441-4514, USA
danb@rl.af.mil

Michael I. Savic
ECSE Department/Speech Research Group
Rensselaer Polytechnic Institute
Troy, NY 12180, USA
savic@ecse.rpi.edu

ABSTRACT

This paper presents a voicing state determination algorithm (VSDA) that is used to simultaneously estimate the voicing state of two speakers present in a segment of co-channel speech. Supervised learning trains a *Bayesian* classifier to predict the voicing states. The possible voicing states are silence, voiced/voiced, voiced/unvoiced, unvoiced/voiced and unvoiced/unvoiced. We have assumed the silent state as a subset of the unvoiced class, except when both speakers are silent. We have chosen a binary tree decision structure. Our feature set is a projection of a 37 dimensional feature vector onto a single dimension applied at each branch of the decision tree, using the Fisher linear discriminant. We have produced co-channel speech from the TIMIT database which is used for training and testing. Preliminary results, at signal to interference ratio of 0 dB, have produced classification accuracy of 82.6%, 73.45%, and 68.24% on male/female, male/male and female/female mixtures respectively

1. INTRODUCTION

Voicing state determination is a method of classifying the voicing state of a segment of speech. We extend this concept to co-channel speech signals, where we now must classify the voicing state of multiple speakers present in a segment of co-channel speech. This process is required in a co-channel speaker separation system as a means to select an appropriate separation processing technique [2]. Other attempts have been made at a VSDA for co-channel speech, but only with limited success using *a priori* information [6]. Similar attempts have also been made in identifying the number of talkers within a given segment [7].

The possible voicing classifications we have considered for co-channel speech are:

1. Silence (SIL) - both speakers are silent;
2. Voiced/Voiced (V/V) - both speakers are producing voiced sounds;
3. Voiced/Unvoiced (V/UV) - the desired speaker is voiced while the undesired speaker is unvoiced;
4. Unvoiced/Voiced (UV/V) - the desired speaker is unvoiced while the undesired speaker is voiced;
5. Unvoiced/Unvoiced (UV/UV) - both speakers are unvoiced.

Classifying co-channel speech requires simultaneously estimating the voicing state of each speaker present within the segment. We have assumed the silent state as a subset of the unvoiced class (except when both speakers are silent) thereby limiting classification of co-channel speech to mixtures of voiced and unvoiced speech and total silence.

In this work we have developed a VSDA based on a decision theory. The detector can be modeled as a black box with a set of inputs and a set of outputs. The box operates in both a training mode and a detection mode. In training mode, the detector is presented with co-channel speech data segments from which it then creates a reference associated with the five classes defined above. Once training is complete, the detector operates in a recognition mode in which it is presented with an unknown set of data. The detector is then tasked to identify which of the five possible voicing classes should be assigned to the data. The detector is evaluated based on its ability to correctly classify unknown co-channel speech segments.

2. DECISION STRUCTURE

Voicing state determination of co-channel speech requires discrimination between five classes of speech. There are several ways in which an R-category classification can be structured. Classification can be obtained by a single classifier which assigns the pattern to one of R classes, or through a sequence of binary decisions. We have chosen a binary decision tree approach to classification.

The binary decision tree structure is shown Figure 1. Decisions are made on a frame-by-frame basis. The first decision is to decide on the presence or absence of speech in a

given speech segment. If the decision is made that no speech is present, then the segment is labeled as *silence*. If speech is present, we move down the decision tree to the next level. Here we must decide if there is voiced speech present or strictly unvoiced speech present. If only unvoiced speech is present, the segment of speech is labeled as *unvoiced/unvoiced*. If voiced speech is present, we proceed down the decision tree to the next branch to determine if both speech signals are voiced or if one signal is voiced and the other is unvoiced. If both signals are voiced, we label the speech as *voiced/voiced*. If the speech segment is combination of voiced and unvoiced speech, we continue down to the last branch to decide which speaker is voiced and which is unvoiced. Here the speech is labeled as *voiced/unvoiced* or *unvoiced/voiced*.

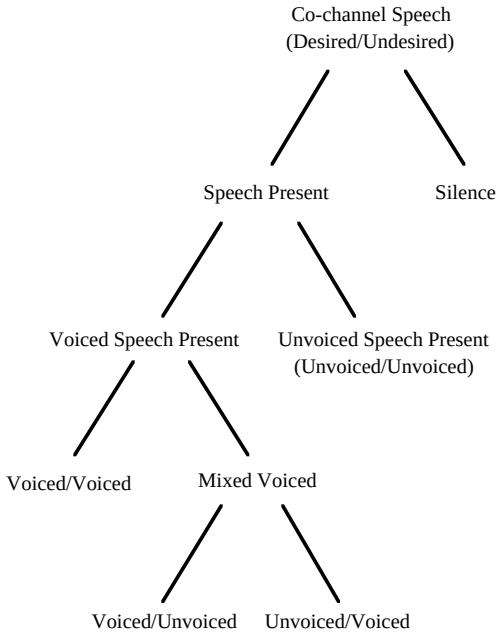

Figure 1: Voicing state decision tree for co-channel speech.

3. FEATURES

The selection of a set of features that will provide adequate classification of co-channel speech must be more sophisticated than those used for voicing state classification of uncorrupted speech. The set of features chosen must not only discriminate between classes of voiced and unvoiced speech, but it must also discriminate between mixed excitation of two speakers. That is, the feature set must successfully discriminate between the sum of two voiced segments of speech from the sum of a voiced and unvoiced speech segments. The feature set must also discriminate between mixed excitation between two different speakers. The features we have chosen are:

1. Log of the short time energy of the signal;
2. Normalized fundamental frequency;
3. Normalized autocorrelation coefficient at unit sample delay;
4. Normalized zero crossing rate;
5. Ratio of energy in the signal above 4 kHz to energy below 4 kHz;
6. 16 mel-cepstral coefficients;
7. 15 modified covariance coefficients excluding the first one.

The features considered here are chosen not only for their ability to discriminate between voiced, unvoiced and mixed speech, but also to differentiate between speakers. The first four features are a subset of the traditional voicing state determination systems. The last two features in the set are unique to our application in the discrimination of voiced/voiced speech from mixed voiced speech and in discriminating mixed voiced speech between speakers.

4. TRAINING DATA

In this work we are developing a pattern recognition approach for deciding the class of speech based on measured features from the co-channel speech signal. The classes of speech are those identified above. Our classifier is trained to recognize patterns of speech through supervised training. Training is accomplished using the TIMIT database. The TIMIT database contains clean English spoken speech sampled at 16 kHz. The database is segmented into eight distinct dialect regions of the United States. We have performed training and testing on Northern USA speakers.

The TIMIT database provides a hand labeled phonetic transcription of each sentence within the database. Logically, this would appear to be the most accurate way to segment the speech. However, since a typical phone will transverse across several or more frames, and a phone contains both voiced and unvoiced speech, we have developed our own segmentation and labeling system for training. The features, described above, are extracted from the labeled data and used to train the classifier.

The system used to segment uncorrupted speech for training our classifier is shown in Figure 2. The short time energy, along with the zero crossing rate are two features that have proven to be effective in making a voiced/unvoiced classification of uncorrupted speech [1].

The energy threshold is defined as

$$E_{thrshld} = .6 * min\left(\frac{1}{M} \sum_{i=1}^{M} 10 * log(E_i) \right) \qquad (1)$$

where E_i is the short time energy measure per frame and M is the total number of frames within the length of the co-channel speech signal. Speech below the energy threshold is classified as silence.

The threshold for zero-crossing rate (ZCR) is predetermined. It is based on the sampling rate and the frame size used in the windowing routine. The threshold for the ZCR is given by [4]

$$ZCR_{Thrhld} = \frac{2480}{f_s} * N \qquad (2)$$

where f_s is the sampling rate and N is the number of samples per frame. When the ZCR is greater than the given threshold and the energy is greater than the threshold, the segment of speech is labeled as unvoiced. If the ZCR is less than, or equal to the threshold, and the energy is greater than the threshold, the speech segment is labeled as voiced. Otherwise the speech segment is labeled as silence

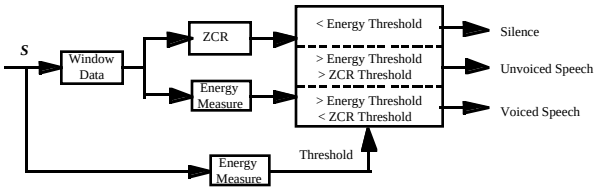

Figure 2: Voiced/Unvoiced segmentation of speech

5. BAYESIAN APPROACH

Based on the decision structure presented above, our problem of identifying the voicing states of speech segments becomes a sequential series of decisions between two classes. In the next two sections, we treat the classification of speech as a two-class problem. In our two class problem, hypothesis H_0 is true when our measurement x belongs to class 0 and hypothesis H_1 is true when x belongs to class 1. Class i will be defined as a termination point in our decision tree.

The Bayes decision rule for minimum error, on a two-class problem is as follows: given an observation vector $\boldsymbol{x}$, the classifier decides hypothesis H_0 if the probability of H_0 is greater than the probability of H_1. Otherwise, the decision is H_1. This can be written as a likelihood ratio test [5]

$$l(\boldsymbol{x}) = \frac{p(\boldsymbol{x}/H_1)}{p(\boldsymbol{x}/H_0)} \underset{H_0}{\overset{H_1}{\gtrless}} \frac{P_0}{P_1} \qquad (3)$$

where the P_i is the *a priori* probability of hypothesis H_i, and $p(\boldsymbol{x}/H_i)$ is the conditional density function. The term P_0/P_1 becomes the threshold value of the likelihood ratio decision. Here we assume that no cost is associated with a wrong decision.

To achieve minimum error rate classification under the Bayes decision rule, we must chose our classification such that it minimizes the conditional risk. Thus, we must decide the hypothesis that maximizes the a posteriori probability $p(H_i/\boldsymbol{x})$. The hypothesis with the largest a posteriori probability insures a minimum error rate.

The type of classifier will be dependent on the conditional density functions $p(\boldsymbol{x}/H_i)$. The likelihood ratio takes on an analytically attractive form when the density function is multivariate normal. A multivariate normal density function is defined as

$$p(\boldsymbol{x}) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} exp\left[-\frac{1}{2}(\boldsymbol{x}-\mu)^t \Sigma^{-1}(\boldsymbol{x}-\mu)\right] \qquad (4)$$

where μ is the n-component mean vector and Σ is the n-by-n covariance matrix. Unfortunately, our n-dimensional vector is not multivariate normal. We can however form a linear combination of the components of $\boldsymbol{x}$ that will project this n-dimensional vector onto a line. We can write this projection as

$$y = \boldsymbol{w}^t \boldsymbol{x} \qquad (5)$$

where y is a linear sum of the elements of $\boldsymbol{x}$. If this transformation is chosen properly, we can project these vectors in such a manner that the samples are well separated.

To insure that the samples are well separated, the distance between the means of the projected samples must be large while maintaining the variances of these projected samples to be small. The Fisher linear discriminant [3] is a linear function y defined in equation (5) such that the criterion function

$$J(\boldsymbol{w}) = \frac{|\tilde{m}_1 - \tilde{m}_2|^2}{\tilde{s}_1 + \tilde{s}_2} \qquad (6)$$

is maximum, where $\tilde{m}_i$ is the sample mean of the projected points for class i, and $\tilde{s}_i$ is the scatter of the projected samples for class i.

If the elements of $\boldsymbol{x}$ are mutually independent, the dimension of $\boldsymbol{x}$ is large and the components of y satisfy the Lindeberg conditions, consequently from the *central limit theorem*, y can be taken to be a normal random variable. The Lindeberg condition states that the individual variances σ_k^2, for k = 1,...,n must be small compared to the sum of all the variances, $\sum_{k=1}^{n} \sigma_k^2$. The assumption that y is a normal random variable provides an optimum partitioning of the real line into two decision regions.

Referring back to equation (3), we now develop a classifier based on the statistical characteristics of our data. We can view the likelihood ratio test in (3) in terms of a set of discriminant functions $g_i(\boldsymbol{x})$ for each hypothesis or class. Our classifier assigns an observation $\boldsymbol{x}$ to the class with the largest discriminant. For the minimum error rate, our discriminant functions becomes

$$g_i(\boldsymbol{x}) = p(H_i/\boldsymbol{x}) \qquad (7)$$

such that the maximum discriminant function is the maximum a posteriori probability. Using the Bayes rule, taking the

logarithm of both sides, and simplifying our expression we can write our discriminant function as

$$g_i(y) = -\frac{1}{2}(y-\mu_i)^t \sigma_i^{-1}(y-\mu_i) - \frac{1}{2}log(|\sigma_i|) + log\,P(H_i) \quad (8)$$

The discriminant functions are quadratic and the decision regions lie along a straight line. This procedure is not optimal. Projection of an *n-dimensional* vector onto a real line can not reduce the minimum achievable error rate. We are throwing away information which may aid in the classification. However, this technique allows the use of the Bayes rule with a univariate normal density function which is mathematically attractive and has the added advantage of working in a single dimension.

6. RESULTS

Three different sets of speech mixtures were created to measure performance of our voicing state determination algorithm. The spoken language was English and the speech signals were taken from the TIMIT database. The co-channel speech mixtures consisted of a male/female, male/male and female/female mixtures. All three mixtures were tested at a desired to interfering speaker ratio (SIR) of 0 dB. We also tested the male/female data set at a SIR = -6 dB.

The data sets for each mixture was comprised of ten overlapping speech sentences using two speakers. Cumulative results of all ten sentence combinations for each speech mixture are presented below. The results along the main diagonal in the tables represent the percentage of co-channel speech segments labeled correctly. The numbers down each column, off the main diagonal, represent the percentage of incorrectly labeled segments belonging to a particular class (missed detection). The numbers along each row, off the main diagonal, represent the percentage of speech segments which were incorrectly identified as belonging to a particular class (false detect). The VSDA performed better on the different sex mixtures than on the same sex mixtures. For the different sex mixtures, the performance of the VSDA improved for SIR = -6 dB compared to the same speech sentence mixtures at SIR = 0 dB. A high percentage of errors occurred either during the onset or offset of voicing of one or both speakers within the interval of speech. Other features and classifiers need to be considered to improve the classification of the co-channel speech segments.

REFERENCES

[1] B.S. Atal and L.R. Rabiner, "A Pattern Recognition Approach to Voiced-Unvoiced-Silence Classification with Applications to Speech Recognition," *IEEE Trans. on Acoust., Speech and Signal Processing*, vol. ASSP-24(3), June 1976.

[2] D. S. Benincasa and M. I. Savic, "Co-channel Speaker Separation using Constrained Nonlinear Optimization," International Conference on Acoustic Speech and Signal Processing, Munich, Germany, April 1997.

[3] R. O. Duda and P.E. Hart, *Pattern Classification and Scene Analysis*, John Wiley & Sons, New York, NY, 1973.

[4] Douglas O'Shaughnessy, *Speech Communication*, Addison-Wesley Publishing Company, Reading MA, 1987.

[5] H. L. Van Trees, *Detection, Estimation, and Modulation Theory Part I*, John Wiley & Sons, New York, NY, 1968.

[6] M. Weintraub, "A Computational Model for Separating Two Simultaneous Talkers," International Conference on Acoustic Speech and Signal Processing, Tokyo, Japan , 1986.

[7] M. A. Zissman, "Cochannel Talker Interference Suppression," Technical Report 895, Lincoln Laboratory, MIT, 26 July 1991.

Table 1: Male/Female mixtures, SIR = 0 dB. Overall 82.60% of the speech segments were correctly identified.

Voicing	SIL	V/V	V/UV	UV/V	UV/UV
SIL	99.55	0	0.51	0.36	5.43
V/V	0	90.14	24.30	14.41	0
V/UV	0	6.38	65.47	2.70	5.16
UV/V	0	3.48	3.84	79.46	8.15
UV/UV	0.45	0	5.88	3.06	81.25

Table 2: Male/Male mixtures, SIR = 0 dB. Overall 73.45% of the speech segments were correctly identified.

Voicing	SIL	V/V	V/UV	UV/V	UV/UV
SIL	100	0	1.07	0.17	21.40
V/V	0	93.33	29.87	38.10	0.42
V/UV	0	3.97	58.93	4.48	5.08
UV/V	0	2.70	5.60	53.10	11.44
UV/UV	0	0	4.53	4.14	61.65

Table 3: Female/Female mixtures, SIR = 0 dB. Overall 68.24% of the speech segments were correctly identified.

Voicing	SIL	V/V	V/UV	UV/V	UV/UV
SIL	100	0	0.32	0.58	10.58
V/V	0	88.87	34.46	31.86	0.53
V/UV	0	4.52	48.24	10.17	7.41
UV/V	0	6.61	14.74	53.74	8.47
UV/UV	0	0	2.24	3.65	73.02

Table 4: Male/Female mixture, SIR = -6 dB. Overall 84.22% of the speech segments were correctly identified.

Voicing	SIL	V/V	V/UV	UV/V	UV/UV
SIL	100	0	0.77	0	5.09
V/V	0	91.75	19.74	10.75	0
V/UV	0	3.91	67.69	3.23	7.38
UV/V	0	4.34	4.36	83.51	6.87
UV/UV	0	0	7.44	2.51	80.66