

Correspondence

Likelihood Decision Boundary Estimation Between HMM Pairs in Speech Recognition

Levent M. Arslan and John H. L. Hansen

Abstract—In maximum likelihood (ML) estimation of hidden Markov models (HMM's) for speech recognition, the criterion is to maximize the total probability across the training data for a particular speech unit, such as a word, monophone, diphone, or triphone. Since each unit model is trained separately, such a strategy can often lead to biases among decision boundaries of the generated model set. In this correspondence, we propose a new technique to minimize the total number of misclassifications in the training data set by adjusting the decision boundaries between HMM pairs. The proposed algorithm is shown to reduce the error rate in a number of speech recognition tasks such as accent detection, language identification, and confusable word pair discrimination. The technique is also attractive because it is simple to implement and the improvement in performance is achieved without any added complexity in the decoding phase.

I. INTRODUCTION

Recent developments in training methods have resulted in speech recognition systems that achieve acceptable error rates under ideal conditions [8], [11]. Advances in language modeling, context-dependent strategies, and speech unit partitioning have increased research efforts in the application of hidden Markov models (HMM's) for speech recognition. The basic theory of HMM's was first introduced by Baum *et al.* more than 20 years ago [5]. The implementation of HMM's for speech processing applications was considered by Baker [4] and later Jelinek *et al.* [9]. Presently, there is no known analytical solution for an HMM that maximizes the probability of a given observation sequence. However, an iterative procedure known as the expectation-maximization algorithm can be employed to estimate the HMM parameters. Unfortunately, this method is not guaranteed to find the optimal solution, and in practice often results in a local maximum. As a result, there are often imperfections in the generated models. These imperfections lead to suboptimal performance of speech recognition systems. One possible solution for this problem is modifying the training algorithm to incorporate the minimization of an empirical function of the recognition error rate over the training data [10]. In this work, we are proposing another solution that adjusts the decision thresholds among model likelihoods in the scoring system to minimize the total number of errors in the training set. Two new methods are described to automatically estimate the optimal decision boundary among confusable speech pattern models. The decision boundaries between HMM likelihoods were used previously in utterance and speaker verification systems [15], [16]. However, to the author's knowledge,

Manuscript received December 7, 1995; revised August 25, 1997. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Mazin Rahim.

L. M. Arslan was with the Robust Speech Processing Laboratory, Department of Electrical Engineering, Duke University, Durham, NC 27708-0291 USA. He is currently with the Entropic Research Laboratory, Washington, D.C. 20001 USA.

J. H. L. Hansen is with the Robust Speech Processing Laboratory, Department of Electrical Engineering, Duke University, Durham, NC 27708-0291 USA (e-mail: jhlh@ee.duke.edu).

Publisher Item Identifier S 1063-6676(98)04220-5.

they have not been utilized before to improve discrimination ability of speech recognition systems.

The outline of this correspondence is as follows. In Section II, we present the motivation and formulation of the proposed likelihood decision boundary estimation method. Two different methods are described in order to estimate the decision boundary between model pairs. Section III presents evaluations of the likelihood decision boundary estimation procedure for the problems of accent detection, pairwise language identification, and confusable word pair discrimination in speech recognition. Finally, Section IV summarizes contributions and draws conclusions.

II. LIKELIHOOD DECISION BOUNDARY ESTIMATION

Many errors that occur in speech recognition system evaluation are due to confusable word pairs in the vocabulary set (e.g., discrimination of "white" versus "wide," or "six" versus "fix"). Some studies suggest that performance can be improved if more distinct vocabulary sets are selected. However, this places an added burden on the speech recognition system designer and may lead to less natural command selection. When the confusion matrices of recognition systems are analyzed, frequently there are unfair biases (i.e., resulting in suboptimal performance) toward one of the words in the confusable pair. Despite this fact, in general the output of Viterbi scoring is taken for granted in speech recognition systems. In this work, we are proposing a method to achieve optimal performance given a fixed HMM set by adjusting the decision boundaries among models properly. It should be noted here that we are not suggesting to force uniform distribution of errors between model pairs. The goal of the method proposed here is to adjust the decision boundary between likelihoods of model pairs with the constraint of *minimizing the overall error rate* in the training set.

A. Theory of Likelihood Decision Boundary Estimation

Normally, when an observation vector sequence \mathbf{X} is to be classified as one of two HMM's λ_1 and λ_2 , the conditional probabilities $P(\mathbf{X}|\lambda_1)$ and $P(\mathbf{X}|\lambda_2)$ are calculated and the model resulting in the higher likelihood is selected (assuming that *a priori* probabilities are equal). Therefore, the decision procedure can be expressed as

$$\begin{aligned} \text{Choose } \lambda_1, & \quad \text{if } P(\mathbf{X}|\lambda_1) > P(\mathbf{X}|\lambda_2) \\ \text{Choose } \lambda_2, & \quad \text{otherwise} \end{aligned} \quad (1)$$

or, if we define a log-likelihood ratio θ as $\ln(P(\mathbf{X}|\lambda_1)/P(\mathbf{X}|\lambda_2))$, then the rule becomes

$$\begin{aligned} \text{Choose } \lambda_1, & \quad \text{if } \theta > 0 \\ \text{Choose } \lambda_2, & \quad \text{otherwise.} \end{aligned} \quad (2)$$

Although this rule is mathematically correct and simple, in practice it may not always result in optimal classification performance. This is especially true for speech recognition systems, since these systems are based on models where the criterion is to maximize the probability over the training tokens of each speech unit. In general, this does not guarantee the minimization of the error rate across the whole vocabulary (i.e., since the model for an isolated word, phoneme, or diphone is trained separately from other models in the set). The proposed solution here is to make the adjustment in the output of

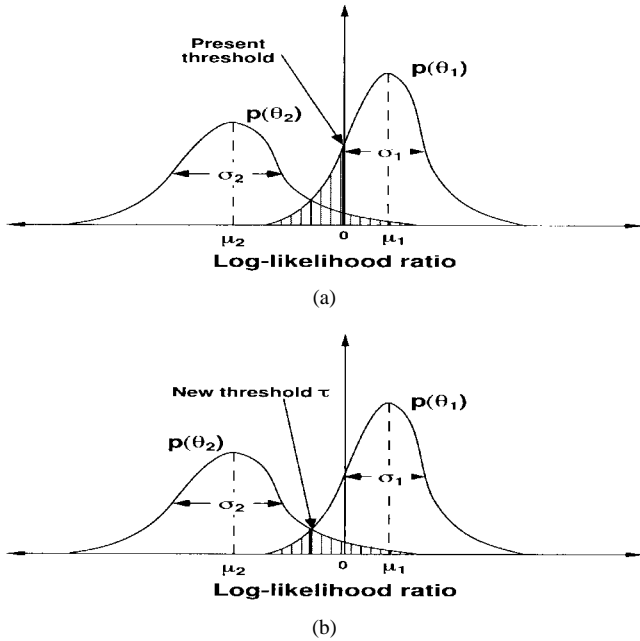


Fig. 1. Illustration of decision boundary estimation between confusable word pairs (a) before and (b) after threshold adjustment.

Viterbi scoring instead of transforming the feature vector (e.g., linear discriminant analysis [6]), or modifying the training procedure to minimize an empirical function of the recognition error rate over the training data [10]. This can be accomplished by updating the threshold, which is normally zero in the decision rule above, which results in a minimization of the total number of errors in the training set. The new rule is expressed as

$$\text{Choose } \lambda_1, \text{ if } \theta > \tau \quad (3)$$

$$\text{choose } \lambda_2, \text{ otherwise} \quad (4)$$

where the threshold τ is determined as follows: let \mathbf{X}_1 and \mathbf{X}_2 denote the training observation vector sets for the confusable word pair $word_1$ and $word_2$, respectively. The statistical models generated from the training set are denoted as λ_1 and λ_2 . Next, we define the following two log-likelihood ratio functions θ_1 and θ_2 , where N_1 is the number of training tokens for $word_1$ and N_2 the number for $word_2$

$$\theta_{1n} = \ln \left(\frac{P(\mathbf{X}_{1n}|\lambda_1)}{P(\mathbf{X}_{1n}|\lambda_2)} \right) \quad n = 1, \dots, N_1 \quad (5)$$

$$\theta_{2n} = \ln \left(\frac{P(\mathbf{X}_{2n}|\lambda_1)}{P(\mathbf{X}_{2n}|\lambda_2)} \right) \quad n = 1, \dots, N_2. \quad (6)$$

In Fig. 1, probability density functions $p(\theta_1)$ and $p(\theta_2)$ are plotted for a typical case where there is a bias toward one of the models. In Fig. 1(a), it can be seen that model 2 is generally selected when erroneous decisions are made. Here, two approaches are suggested to estimate the decision boundary depending on the number of classification errors in the training data.

B. Optimum Decision Boundary Search

For the first method, the procedure is applicable if the number of classification errors in the training set is statistically significant. An incremental search for the decision boundary τ can be performed with the condition of minimizing the total number of errors in the training set. The flowchart for the search procedure is shown in Fig. 2. Here, the log-likelihood values θ_1 and θ_2 for each training token are first generated. Next, an incremental search within the

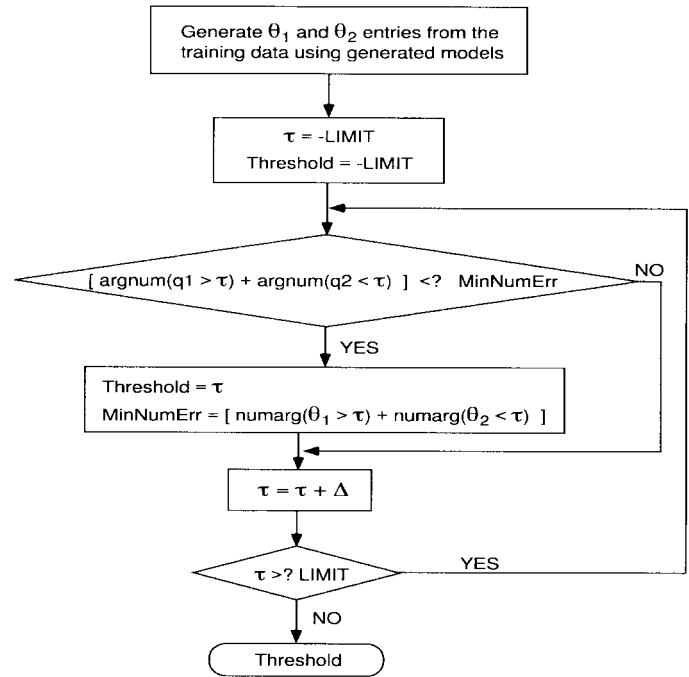


Fig. 2. Flowchart for optimum decision boundary search.

$-LIMIT$ to $+LIMIT$ range specified for the optimum threshold value is performed. This procedure finds a threshold that results in the minimum combined number of errors for the confusable word pair training tokens.

When few errors exist in the training set, this method can still be applied. In such cases, near misses within a certain range could also be regarded as errors in order to have a sufficient number of errors to estimate the decision boundary.

C. Bayesian Decision Boundary Estimation

In the second method, the procedure is more suitable if the number of classification errors in the training set is small and not sufficient to make a reliable threshold estimation using the optimum decision boundary search method. In such a case, estimates of the probability density functions $p(\theta_1)$ and $p(\theta_2)$ can be computed based on a Gaussian assumption. In Fig. 1, the estimate of the new threshold τ is shown as the point where the two distributions intersect. This threshold value minimizes the size of the shaded area, which is the total error region assuming equal *a priori* probabilities for the two models. If the two probability density functions are defined as follows, with given means μ_i and variances σ_i^2 , $p(\theta_1) = N(\mu_1, \sigma_1^2)$, and $p(\theta_2) = N(\mu_2, \sigma_2^2)$, then the decision boundary τ can be computed as follows:

$$\frac{1}{\sqrt{2\pi}\sigma_1} e^{-((\tau-\mu_2)^2/2\sigma_1^2)} = \frac{1}{\sqrt{2\pi}\sigma_2} e^{-((\tau-\mu_2)^2/2\sigma_2^2)} \quad (7)$$

which reduces to

$$\frac{\sigma_2}{\sigma_1} = \frac{e^{-((\tau-\mu_2)^2/2\sigma_2^2)}}{e^{-((\tau-\mu_1)^2/2\sigma_1^2)}}. \quad (8)$$

If the natural logarithm is applied to both sides, [8] reduces to

$$\ln(\sigma_2) - \ln(\sigma_1) = -\frac{(\tau - \mu_2)^2}{2\sigma_2^2} + \frac{(\tau - \mu_1)^2}{2\sigma_1^2}. \quad (9)$$

If these terms are rearranged, the equation reduces to a second-order polynomial of the form:

$$A\tau^2 + B\tau + C = 0, \quad (10)$$

TABLE I
ACCENT DETECTION PERFORMANCE IMPROVEMENT BY DECISION BOUNDARY ADJUSTMENT PROCEDURE OVER CLOSED AND TEST SETS

<i>Error rates in classification of neutral versus accented words</i>				
<i>Test set</i>	<i>Closed Set</i>		<i>Open Set</i>	
	<i>Viterbi</i>	<i>After Threshold Adjustment</i>	<i>Viterbi</i>	<i>After Threshold Adjustment</i>
(Neutral-Turkish)	2.54	1.24	24.75	22.67
(Neutral-Chinese)	1.91	1.04	21.09	18.95
(Neutral-German)	6.11	3.55	26.77	20.67
Overall	3.49	1.93	23.99	20.85

TABLE II
NUMBER OF SPEAKERS FROM EACH LANGUAGE GROUP FOR THE INITIAL TRAINING AND DEVELOPMENT TEST SETS IN OGI MULTILINGUE DATA BASE

<i>OGI MULTI-LANGUAGE SPEECH CORPUS</i>				
<i>Language</i>	<i>Initial Training</i>		<i>Development Test</i>	
	<i>Male</i>	<i>Female</i>	<i>Male</i>	<i>Female</i>
English	33	17	14	6
Farsi	39	10	15	4
French	40	10	15	5
German	25	25	11	9
Korean	47	3	13	4
Japanese	30	20	15	5
Mandarin	34	15	14	6
Spanish	34	16	16	4
Tamil	43	7	17	3
Vietnamese	31	19	16	4

which has the solution

$$\tau_{1,2} = \frac{-B \pm \sqrt{B^2 - 4AC}}{2A} \quad (11)$$

where

$$\begin{aligned} A &= \sigma_1^2 - \sigma_2^2, \\ B &= 2\sigma_2^2\mu_1 - 2\sigma_1^2\mu_2, \\ C &= \sigma_1^2\mu_2^2 - \sigma_2^2\mu_1^2 + 2\sigma_1^2\sigma_2^2(\ln(\sigma_2) - \ln(\sigma_1)). \end{aligned} \quad (12)$$

It is clear that the coefficients are dependent on the means and variances of the two distributions. The solution τ , which lies between the means μ_1 and μ_2 , is chosen as the new decision boundary. It should be noted here that the design criterion for Bayesian decision boundary estimation is set for equal error rate, which does not assure minimum error rate as in the optimum decision boundary search. However, the Bayesian method is proposed as an alternative to the previous method for cases where there are not statistically significant numbers of misclassifications in the training set.

III. EVALUATIONS

For the purpose of evaluating the proposed method, we conducted three experiments in the areas of accent detection, speech recognition (between confusable speech units), and pairwise language identification (ID) using the OGI multilanguage data base.

A. Evaluations on Accent Detection

The optimum decision boundary search method was applied to foreign accent detection. The method was used to distinguish between neutral American and a second accent class using a data base collected at the Duke University Robust Speech Processing Laboratory. A vocabulary of isolated words and phrases was established, which contains accent sensitive phonemes or phoneme combinations [1],

[3]. Vocabulary choice was based on a literature review of language education of American English as a second language. A portion of the data corpus was collected using a head-mounted microphone in a quiet office environment, while the remaining part was collected through an on-line telephone interface at an 8 kHz sampling rate. All speakers were from the general Duke University community, with 43 speakers using microphone data entry and 68 speakers using on-line telephone data entry. The test vocabulary consists of 20 accent sensitive isolated words such as "aluminum," "thirty," "bringing," "target," etc., and four American English phrases. Every speaker repeated each word five times. The accent data base includes neutral American English, and English under the following accents: German, Chinese, Turkish, French, Persian, Spanish, Italian, Japanese, and others. The studies conducted here focused on American English speech from 76 speakers across the following accents: neutral, Turkish, Chinese, and German (the head-mounted microphone data was bandpass filtered between 100 and 3800 Hz to provide spectral match with data collected through the on-line telephone interface).

For each word in the vocabulary, a continuous mixture HMM with two Gaussian mixtures per state was generated for each accent type. The number of states assigned to each word was proportional to the overall word duration. The number of states typically ranged from 7–21, and all speech data were parameterized using mel-frequency cepstrum and delta coefficients, energy and delta energy. Nine speakers from each accent class were used for training, and a total of 40 speakers were set aside for open set testing. Using the forward-backward training algorithm, a total of 80 HMM's (20 words under four accent types) were generated. A total of 3250 word tokens was used for training, and 3900 word tokens were set aside for open testing. The standard Viterbi algorithm without decision boundary adjustment resulted in an error rate of 3.49% in the closed set where a decision was made between the neutral English model and another accent model (i.e., either German, Chinese, or Turkish). Using the estimated decision boundary between the accent model pair (neutral-versus-other) for each word, the error rate was reduced to 1.93% for the closed set (a 44.7% reduction from original). For the open test set, the error rate was reduced from 23.99 to 20.85% (a 13.1% reduction with detailed results summarized in Table I). While there is measurable error rate reduction for both closed and open test sets, it is noted that the level of performance improvement was more for closed versus open data. It is believed that this occurs because the optimum boundary decision method concentrates highly on minimizing the error rate of the training data.

B. Evaluations Using the SUSAS Data Base

Next, the proposed decision boundary estimation method is applied to a series of confusable word pairs from the SUSAS speech-under-stress data base [7]¹. A common, highly confusable, vocabulary set of 35 aircraft communication words consisting of mono- and

¹Approximately half of the SUSAS data base consists of style data donated by the MIT Lincoln Laboratory [12].

TABLE III
LANGUAGE ID PERFORMANCE IMPROVEMENT AFTER THRESHOLD ADJUSTMENT FOR PAIRWISE ENGLISH-OTHER EXPERIMENTS

<i>Error rates in classification of English versus other languages in OGI multi-language database</i>				
<i>Test set</i>	Closed Set		Open Set	
	<i>Viterbi</i>	<i>After Threshold Adjustment</i>	<i>Viterbi</i>	<i>After Threshold Adjustment</i>
English-Farsi	7.8	6.6	16.2	13.6
English-French	9.7	7.7	15.8	18.3
English-German	10.0	8.9	29.7	24.6
English-Japanese	4.4	4.4	8.6	8.3
English-Spanish	9.9	6.4	21.6	16.2
English-Korean	4.8	5.8	8.3	8.3
English-Mandarin	0.0	0.0	8.1	8.1
English-Tamil	3.4	2.3	16.2	11.0
English-Vietnamese	2.4	3.5	16.7	16.7
Overall	5.8	5.1	15.7	13.9

multisyllabic words make up the data base. A more complete discussion of SUSAS can be found in the literature [7]. SUSAS data used in this study consist of nine adult male speaker utterances, all sampled at 8 kHz using a 16-b A/D converter. Nine of the most confusable word pairs in the data base were selected as test material for the proposed method (“white-wide,” “six-fix,” “oh-no,” “oh-go,” “go-no,” “east-eight,” “east-eighty,” “change-gain,” “east-degree”). Continuous density HMM’s were generated for each word considered as being confusable. A decision boundary was estimated for each word model pair from the speech of six speakers (12 tokens each). In the test set, two different speakers and ten different stress conditions from all speakers in the data base were included. The stress conditions include speech that is slow, fast, soft, loud, angry, clear, question, and Lombard effect (i.e., speech spoken in noise). A total of 3024 confusable word tokens were used in the open test set evaluation. The error rate was found to be 5.89% when the Viterbi algorithm was used with a zero threshold for the likelihood ratio. When the Bayesian decision boundary estimation was used to shift the thresholds for the confusable word pairs, the error rate reduced to 5.16% (a 12.36% reduction).

C. Evaluations on Language ID

The Oregon Graduate Institute Multilanguage Telephone Speech (OGI-TS) Corpus [14] was used to evaluate the performance of decision boundary adjustment method for the problem of language identification. Since the available training data set was sufficiently large, the optimum decision boundary search method was employed. Each message in the corpus was spoken by a unique speaker over a telephone channel and includes responses to ten prompts, four of which contain fixed text (e.g., “Please recite the seven days of the week,” “Please say the numbers zero through ten”) and six of which assume free text responses (e.g., “Describe the room from which you are calling,” “Speak about any topic of your choice”). All together the ten responses contained in each session comprise about two minutes of speech. Table II lists the number of messages per language in each of the two segments of the corpus: initial training, development test.² In our evaluations the initial training set (about 50 speakers per language) was used in training, and development test (about 20 speakers per language) was used in testing. Test utterances were extracted from the development test set according to the April 1993 NIST specification [13]:

²The extended training and final test sets are not considered in this study.

1) “45 s” Utterance Testing: Language ID is performed on a set of 45 s utterances spoken by the development test speakers. These utterances are the first 45 s of the responses to the prompt “speak about any topic of your choice.”

OGI refers to these utterances as “stories before the tone,” and they are denoted *story-bt*.³

In the language ID system developed here, a Gaussian mixture model with 64 mixtures is employed for each language. The system feature set was comprised of 8 cepstrum and 8 delta coefficients plus delta energy. The cepstrum coefficients are computed based on a new accent/language sensitive frequency scale instead of the commonly used mel-scale, which is shown to result in improved performance for the problem of language ID [2]. Pairwise language ID experiments were conducted where the decision was made between English and one of the other nine languages based on 45-s utterances. The results comparing the performance before and after employing the decision boundary adjustment are summarized in Table III. Overall, the error rate was reduced from 15.7% to 13.9% (a 12.4% reduction) after employing decision boundary adjustment method between language model pairs.

IV. SUMMARY AND CONCLUSIONS

In this study, a new technique has been proposed to estimate optimal likelihood decision boundaries between HMM pairs for applications in speech recognition. Two strategies were proposed, which include the optimum decision boundary search and Bayesian decision boundary estimation depending on the number of classification errors in the training set needed to form a reliable threshold estimate. A search procedure was established to obtain the optimum decision boundary, and a closed-form solution was also determined for calculation of the Bayesian decision boundary. The choice of using the optimum decision boundary search versus the Bayesian decision boundary method rests in the amount of training data available, and the resulting error characteristics for the HMM speech recognition pairs under test. When there is a sufficient number of training tokens, then the ability to characterize the output error characteristics is improved, and the optimum decision boundary search method should be employed. When the training data is limited, a sufficient number of recognition errors may not exist in order to estimate the optimal decision boundary. Therefore, under these conditions, the Bayesian decision boundary estimation is recommended. The

³A tone signaled the speaker when 45 s of speech had been collected indicating 15 s remaining.

decision boundary adjustment method was evaluated for three speech recognition applications. First, the proposed method was tested on speech from a foreign accent data base in order to discriminate between neutral and foreign accented speech. Here, American English produced under neutral, German, Chinese, and Turkish accents were considered. In this case, a 13.1% error rate reduction was observed on the open test set. Next, the method was tested on the SUSAS speech-under-stress data base in order to improve discrimination performance between confusable word pairs. For this scenario, an average 12.4% reduction in the error rate was achieved. Finally, the method was evaluated using the OGI multilanguage data base to improve discrimination ability between language pairs. In accordance with the previous results, a reduction of 12.4% in error rate was achieved using the proposed method. It should be noted that the improvements are achieved with no added complexity to the existing systems, which represents an important advantage in using the proposed technique for real-time speech recognition applications. Finally, though the proposed method was formulated and tested under isolated-word recognition scenarios, it can be easily adapted to more general problems in continuous speech recognition and other speech or speaker classification problems. In particular, the proposed methods for automatic estimation of the optimal decision boundary can be utilized in utterance verification systems, where the decision boundaries are estimated based on heuristic methods or adjusted manually in most systems.

REFERENCES

- [1] L. M. Arslan and J. H. L. Hansen, "Language accent classification in American English," *Speech Commun.*, vol. 18, pp. 353–367, Aug. 1996
- [2] —, "A study of temporal features and frequency characteristics in American English foreign accent," *J. Acoust. Soc. Amer.*, vol. 102, pp. 28–40, July 1997.
- [3] —, "Foreign accent classification using source generator based prosodic features," in *Proc. IEEE ICASSP*, Detroit, MI, 1995, pp. 836–839.
- [4] J. K. Baker, "The Dragon system: An overview," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-23, pp. 24–29, 1975.
- [5] L. E. Baum and J. A. Eagon, "An inequality with applications to statistical estimation for probabilistic functions of Markov processes and to a model for ecology," *Bull. Amer. Math. Soc.*, vol. 73, pp. 360–363, 1967.
- [6] R. A. Fisher, "The use of multiple measures in taxonomic problems," *Contr. Math. Stats.*, 1950, pp. 32.179–32.188.
- [7] J. H. L. Hansen, "Morphological constrained enhancement with adaptive cepstral compensation for speech recognition in noise and Lombard effect," *IEEE Trans. Speech Audio Processing*, vol. 2, pp. 598–614, Oct. 1994.
- [8] X. Huang *et al.*, "Microsoft windows highly intelligent speech recognizer: WHISPER," in *Proc. IEEE ICASSP*, Detroit, MI, 1995, pp. 93–96.
- [9] F. Jelinek, "Continuous speech recognition by statistical methods," *Proc. IEEE*, vol. 64, pp. 532–556, 1976.
- [10] B. H. Juang and S. Katagiri, "Discriminative learning for minimum error classification," *IEEE Trans. Signal Processing*, vol. 40, pp. 3043–3054, 1992.
- [11] K. F. Lee, *Automatic Speech Recognition: Development of SPHINX System*. Boston, MA: Kluwer, 1989.
- [12] R. P. Lippmann, E. A. Martin, D. B. Paul, "Multi-style training for robust isolated-word speech recognition," in *Proc. IEEE ICASSP*, Dallas, TX, Apr. 1987, pp. 705–708.
- [13] A. F. Martin, "Language ID guidelines and results," Technical report, Spoken Lang. Proc. Grp., Nat. Inst. Stand. Technol., Gaithersburg, MD.
- [14] Y. K. Muthusamy, R. A. Cole, and B. T. Oshika, "The OGI multilanguage telephone speech corpus," in *Proc. Int. Conf. Spoken Language Processing*, Oct. 1992. vol. 2, pp. 895–898.
- [15] A. E. Rosenberg, C. H. Lee, B.-H. Juang, and F. K. Soong, "The use of cohort normalized scores for speaker verification," in *Proc. Int. Conf. Spoken Language Processing*, Oct. 1992, vol. 2, pp. 599–602.
- [16] R. A. Sukkar and C.-H. Lee, "Vocabulary independent discriminative utterance verification for nonkeyword rejection," *IEEE Trans. Speech Audio Processing*, vol. 4, pp. 420–429, Nov. 1996.

Technique for Broadband Correlated Interference Rejection in Microphone Arrays

Darren B. Ward

Abstract—A new technique for broadband minimum variance beamforming is presented that overcomes the signal cancellation problem of conventional adaptive beamformers in the presence of correlated interference. Specifically, frequency-invariant beamforming is used to perform frequency domain averaging, thereby reducing the correlation between the desired signal and the interference. Such a technique is useful for hands-free speech acquisition using a microphone array, since correlated interference will be present due to room reflections.

Index Terms—Adaptive beamforming, microphone arrays, signal cancellation.

I. INTRODUCTION

Conventional adaptive beamforming methods exhibit cancellation of the desired signal when employed in environments containing interference that is correlated with the desired signal [1]. In trying to achieve its goal of minimum output power, the beamformer uses the correlated interference to cancel part of the desired signal. This problem arises in hands-free speech acquisition using a microphone array, since reflections from walls and other surfaces produce interference that is highly correlated with the desired speech signal.

Methods of dealing with correlated interference generally rely on either spatial averaging [2] or (for broadband signals) spectral averaging [3], [4] to destroy the correlation. It has also been observed [5] that the signal cancellation problem in microphone arrays may be solved by using sufficiently short transversal filters on the array channels such that the direct and reflected speech signals are not simultaneously present in the beamformer.

In this correspondence, a new optimum beamformer is proposed that uses frequency-invariant beamforming [6] to perform spectral averaging. Unlike the method of [3], no preliminary estimation of the interference directions is required.

II. MINIMUM VARIANCE BEAMFORMING

Consider D broadband farfield source signals impinging on a linear array of N sensors from directions $\Theta = [\theta_1, \dots, \theta_D]$, measured relative to the array axis. One of these signals is a desired signal that

Manuscript received May 15, 1997; revised October 20, 1997. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Dennis R. Morgan.

The author was with the Telecommunications Engineering Group, Australian National University, Canberra, ACT 0200, Australia. He is now with the Acoustics and Speech Research Department, Bell Laboratories, Lucent Technologies, Murray Hill, NJ 07974 USA (e-mail: dbwr@research.bell-labs.com).

Publisher Item Identifier S 1063-6676(98)04216-3.