Advances in Unsupervised Audio Classification and Segmentation for the Broadcast News and NGSW Corpora

Rongqing Huang, *Student Member, IEEE,* and John H. L. Hansen, *Senior Member, IEEE*

***Abstract*—The problem of unsupervised audio classification and segmentation continues to be a challenging research problem which significantly impacts automatic speech recognition (ASR) and spoken document retrieval (SDR) performance. This paper addresses novel advances in 1) audio classification for speech recognition and 2) audio segmentation for unsupervised multispeaker change detection. A new algorithm is proposed for audio classification, which is based on weighted GMM Networks (WGN). Two new extended-time features: variance of the spectrum flux (VSF) and variance of the zero-crossing rate (VZCR) are used to preclassify the audio and supply weights to the output probabilities of the GMM networks. The classification is then implemented using weighted GMM networks. Since historically there have been no features specifically designed for audio segmentation, we evaluate 16 potential features including three new proposed features: perceptual minimum variance distortionless response (PMVDR), smoothed zero-crossing rate (SZCR), and filterbank log energy coefficients (FBLC) in 14 noisy environments to determine the best robust features on the average across these conditions. Next, a new distance metric, T^2-mean, is proposed which is intended to improve segmentation for short segment turns (i.e., 1–5 s). A new false alarm compensation procedure is implemented, which can compensate the false alarm rate significantly with little cost to the miss rate. Evaluations on a standard data set—Defense Advanced Research Projects Agency (DARPA) Hub4 Broadcast News 1997 evaluation data—show that the WGN classification algorithm achieves over a 50% improvement versus the GMM network baseline algorithm, and the proposed compound segmentation algorithm achieves 23%-10% improvement in all metrics versus the baseline Mel-frequency cepstral coefficients (MFCC) and traditional Bayesian information criterion (BIC) algorithm. The new classification and segmentation algorithms also obtain very satisfactory results on the more diverse and challenging National Gallery of the Spoken Word (NGSW) corpus.**

***Index Terms*—Audio classification, audio segmentation, Bayesian information criterion, broadcast news transcription, feature analysis, feature processing, Gaussian mixture model (GMM) networks, noisy environments, rich transcription, speaker segmentation, spoken document retrieval.**

Manuscript received January 30, 2004; revised May 23, 2005. This work was supported in part by the National Science Foundation (NSF) under Cooperative Agreement IIS-9817485. Any opinions, findings, and conclusions expressed in this material are those of the authors and do not necessarily reflect the views of NSF. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Tanja Schultz.

The authors were with the Robust Speech Processing Group, CSLR, University of Colorado, Boulder, CO 80309 USA. They are now with the Center for Robust Speech Systems, University of Texas at Dallas, Richardson, TX 75083-0688 USA (e-mail: huangr@colorado.edu; John.Hansen@utdallas.edu).

Digital Object Identifier 10.1109/TSA.2005.858057

I. Introduction

THE GOAL of audio segmentation and classification is to partition and label an input audio stream into speech, music, commercials, environmental background noise, or other acoustic conditions. This preliminary stage is necessary for effective large vocabulary continuous speech recognition (LVCSR), audio content analysis and understanding, audio information retrieval, audio transcription, audio clustering, and other audio recognition and indexing applications.

While much work has been conducted in segmentation and classification, most studies have considered one homogeneous style of audio data. Audio streams for content analysis generally consist of a variety of materials and formats consistent with the following domains.

Monologues: Single speaker talking spontaneously or reading prepared or prompted text in clean conditions.

Two-Way Conversations: Telephone conversations between two subjects that are spontaneous and could contain periods with both talking.

Speeches: Audio data where a person (e.g., politician) is speaking to an audience, primarily one talker, but background audience noise could be present, and room echo or noise is possible.

Interviews/Debates: Audio streams where a person is being interviewed by a TV or radio person. Debates could include a moderator and/or various audience participation (e.g., questions, applause, interruptions, etc.).

Radio/TV News Broadcasts: Would include traditional news anchor with periods of both prompted read speech, spontaneous speech, background music, commercials, other background audio content (e.g., office noise such as typewriter, etc.). Audio content would come from TV or radio studio settings.

Field News Broadcasts: Audio content would come from news reporters in the field (e.g., emergency or war locations, city streets, etc.), would contain a wide range of background noise content of unpredictable origin. Communication channels would also impact frequency content of the audio.

Recording Media/Transmission: Here, the audio properties can be transformed based on the type of a recording equipment used (e.g., microphones, Edison cylinder disks, reel-to-reel tape, cassette tape, DAT, CD, etc.) or transmission (e.g., AM, FM, voice compression methods—CELP, MELP, ADPCM, etc.).

The purpose for identifying the range of audio stream possibilities here is to emphasize the broad range of speaker and environmental factors that can influence acoustic properties. Most studies in audio classification and segmentation have considered more recent audio content such as the Defense Advanced Research Projects Agency (DARPA) Broadcast News (BN) corpus with materials from the 1990s. However, it is possible that for applications such as digital voice libraries (e.g., NGSW [25], [31], [40]) to contain a much wider range of audio format and content such as that listed above. Our objective for audio classification and segmentation is to help improve speech recognition performance using a LVCSR system for a diverse corpus such as NGSW. Therefore, the audio segmentation should be conducted at the speaker and acoustic condition level and the audio streams would be classified into speech or nonspeech, broadband or narrow band, female or male classes. In this study, we propose to 1) focus on effective feature processing for classification and segmentation and 2) develop integrated classification and segmentation schemes that can be more successful in spontaneous audio streams from a range of acoustic conditions or environmental noise events.

The remainder of this paper is organized as follows. First, background material is presented to help motivate the formulation of new classification and segmentation algorithms. Next, the audio classification algorithm, termed WGN, is proposed in Section III. In Section IV, the compound segmentation (CompSeg) algorithm is introduced. The evaluations are presented in Section V. Finally, a summary of this study is presented.

II. Algorithm Motivation

In this section, background material is presented to help motivate the formulation of the audio classification and segmentation algorithms.

Most audio classification techniques focus on two different aspects: one is the particular feature employed, the other is the statistical model used. Speech and nonspeech (music, songs, environmental sounds, etc.) segments have different distribution characteristics in both the time and frequency domains, so feature-based classification is generally an effective method. A number of studies have focused on alternative feature selection or development. For example, Zhang and Kuo [38] considered the energy function, average zero-crossing rate (ZCR), fundamental frequency and spectral peak tracks as their features; Lu *et al.* [22] considered the noise frame ratio, low short-time energy ratio and four other features; while Li [21] employed the total spectrum power, subband power and also four additional features. The success of feature-based methods depends mainly on the discriminative power of the features, and the methods are implemented either in a complex threshold-dependent scheme [22], [38] or with some pattern classification method (Euclidean distance, nearest neighbor [35], nearest feature line [21], etc.). Model-based classification methods have also been popular recently. Hain *et al.* [12] trained four Gaussian mixture models (GMMs) to classify the DARPA BN data into broadband speech, narrowband speech, speech with music backgrounds, and only music. Ajmera and McCowan [2] compared the GMM with a multilayer perceptron (MLP) and reported that they were comparable; Scheirer and Slaney [28] compared the GMM with maximum *a posteriori* (MAP), K nearest neighbor (KNN), and K-dimension spatial classification (K-d models) and reported that KNN was slightly better than the other three classifiers. The distinction between feature-based methods and model-based methods can be obscure, since many researchers consider both aspects to obtain the best performance improvement. For example, Ajmera and McCowan [2] applied two posterior probability based features: entropy and dynamism for GMM and MLP classifiers; while Scheirer and Slaney [28] compared thirteen features with four classifiers.

The features used in feature-based methods can be considered *extended-time* features, and are represented in the time domain (ZCR, energy, etc.), or in the frequency domain (subband power, low short-time energy ratio, etc.), and are typically not suitable for training a statistical model, especially with a diagonal covariance based GMM. As an alternative, *short-time* features such as the spectral-based MFCC and perceptual minimum variance distortionless response (PMVDR) [37] features are decorrelated and highly independent across the feature vector, and, therefore, suitable for training a statistical model. However, short-time features such as MFCCs encode phoneme level information, which can be inappropriate for speech/nonspeech classification. Based on concepts discussed so far, features used in the feature-based methods and features used in the model-based methods are quite different. As such, most researchers treat the feature-based methods and model-based methods separately and do not consider them in an integrated manner.

In this paper, a novel classification algorithm is proposed that combines the feature and model in a compact way that results in very effective audio classification. First, two new extended-time features, variance of the spectrum flux (VSF), and variance of the ZCR (VZCR) are proposed; next, they are applied for preclassification of input audio streams. The preclassification will produce weights which are supplied to the output probabilities of a GMM network, and then the final classification is implemented using *Weighted GMM Networks (WGN)*. This algorithm combines the feature-based method and model-based method in a compact way rather than in a separate way, and achieves very satisfactory results for audio streams from a range of acoustic scenarios.

The goals of effective audio/speaker segmentation are different than those for automatic speech recognition (ASR), and, therefore, features, processing methods, and modeling concepts successful for ASR may not necessarily be appropriate for segmentation. Features used for speech recognition attempt to minimize the differences across speakers and acoustic environments (i.e., *Speaker Variance*), and maximize the differences across the phoneme space (i.e., *Phoneme Variance*). However, in speaker segmentation for audio streams, it is preferable to maximize speaker traits and minimize the phoneme variances simultaneously to produce segments that contain a single acoustic event or speaker. The traditional MFCC features used for ASR may, therefore, not be as effective for speaker segmentation. Other studies have considered alternative features. For example, Adami *et al.* [1] considered line spectral pair

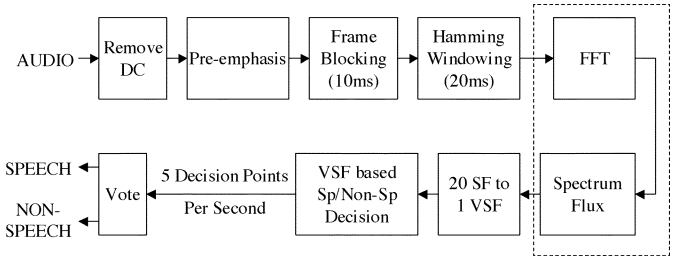


Fig. 1. Process of speech/nonspeech classification with VSF.

(LSP) features, Lu and Zhang [23] used a multifeature set that consisted of the MFCC, LSP, and pitch features to detect change points, and then applied the Bayesian fusion model to combine segmentation results. In the present study, a range of previously developed speech features along with several new features [e.g., PMVDR [37], SZCR, filterbank log energy coefficients (FBLC)] are considered. These features are evaluated in different noise backgrounds in an effort to determine the best feature for segmentation in adverse noisy environments.

If speaker segments are longer than 5 s, the Bayesian information criterion (BIC) [5] and many distance measure-based approaches can achieve reliable segmentation performance [16]. However, these methods suffer from insufficient model estimation traits when the segment turns are short (i.e., less than 5 s). We propose here to use a new distance metric, the T^2-mean, to address this problem. A novel false alarm compensation routine is also developed in the segmentation scheme which can compensate the false alarm rate significantly with little cost to changes in the miss rate. The algorithm is a compound segmentation method, so the scheme is referred to as *CompSeg*.

III. Audio Classification in a Weighted GMM Network

Previous studies have proposed many extended-time features, with the majority derived from the time or frequency domains. Most extended-time features are essentially the same since they encode broad characteristics within the audio. Here, two extended-time features are designed. One is in the time domain, the other is in the frequency domain. Features from different domains might better reflect diverse aspects of the audio structure.

A. VSF

The first feature is the VSF. Fig. 1 shows a flow diagram of the VSF feature extraction process for speech/nonspeech classification.

The spectrum flux (SF) [11], [33] is the ordinary Euclidean norm of the delta spectrum magnitude, which is calculated as

$$SF = \|\mathbf{S_i} - \mathbf{S_{i-1}}\|_2 = \frac{1}{N}\left(\sum_{k=0}^{N-1}\left(S_i(k) - S_{i-1}(k)\right)^2\right)^{\frac{1}{2}} \quad (1)$$

where $\mathbf{S_i}$ is the spectrum magnitude vector of frame i, which is defined as

$$S_i(k) = \left|\sum_{n=0}^{N-1} s\left(n + \frac{Ni}{2}\right) w(n) \exp\left\{\frac{-2\pi kn}{N}\right\}\right|, \quad k \in [0, N-1] \quad (2)$$

where $s(n + (Ni/2))$ is the audio data, N is the size of the window, and $w(n)$ is the window function (i.e., a 20-ms Hamming window is used as shown in Fig. 1).

Actually, the SF itself cannot reflect major differences between speech and nonspeech. It is observed that speech alternates between transient and nonperiodic speech to short-time stationary and periodic speech due to the phoneme transitions (e.g., consonant to vowel, and other phone class transitions). On the other hand, music and environmental sounds could be periodic or monotonic and have more constant rates of change versus that seen in speech. This means the *variance* of SF of speech should be larger than that for music or most environmental sounds. To explore this idea, the following experiment is designed. A set of eight speech clips and eight music clips are selected, where each is 5 s in duration. They are concatenated in an alternating fashion (speech, music, speech, music,. . .). At the end of this 80-s stream, a 10-s speech clip is concatenated followed by a 19-s environmental sound clip. Therefore, the resulting audio stream is 109 s long in duration. Using this 109-s stream, the SF and VSF are calculated, then speech/nonspeech classification is performed using the VSF feature. The SF is calculated on a frame-basis, where the frame size is 20 ms and the frame boundaries are advanced by 10 ms per frame (i.e., 100 frames per second). The VSF is calculated as the variance of the SF over 20 frames. Therefore, in one second of audio, there will be five subblocks, each resulting in a speech/nonspeech decision task as follows:

$$c_{ij} = \begin{cases} 1, & \text{if } V_{ij} \geq T, \\ 0, & \text{else} \end{cases} \quad i = 1, 2, \ldots,\ j = [1, 5] \quad (3)$$

where 1 means speech, 0 means nonspeech, V_{ij} is the VSF value of jth subblock in ith 1-s audio block and T is a threshold. The final decision on the 1-s audio block is based on the vote

$$C_i = \begin{cases} 1, & \text{if } \sum_{j=1}^{5} c_{ij} \geq 3 \\ 0, & \text{else} \end{cases} \quad i = 1, 2, \ldots \quad (4)$$

Fig. 1 shows this decision process. Fig. 2 shows the classification results for the 109-s audio stream. The results show that only 2 s of music was mislabeled as speech from the 109-s passage.[1]

B. VZCR

The second feature is based on the zero-crossing rate (ZCR), which is a commonly used extended-time feature in classification. ZCR is the number of zero-crossings (number of times the sequence changes sign) within a frame in the time domain [18], [27] which is calculated as [6], [26]

$$Z_s(m) = \frac{1}{N}\sum_{n=m-N+1}^{m} \frac{\left|\text{sgn}\{s(n)\} - \text{sgn}\{s(n-1)\}\right|}{2} w(m-n) \quad (5)$$

where N is the length of the frame, m is the endpoint of the frame, and $w(n)$ is the window function. Since music and environmental sounds are more periodic or monotonic than speech, their ZCR will be more constant with less fluctuations. This should mean

[1]The classification processing window in this study is 1 s in duration; however, the evaluation duration unit is 1 frame (i.e., 0.01 s) for comparison with traditional GMM network classification.

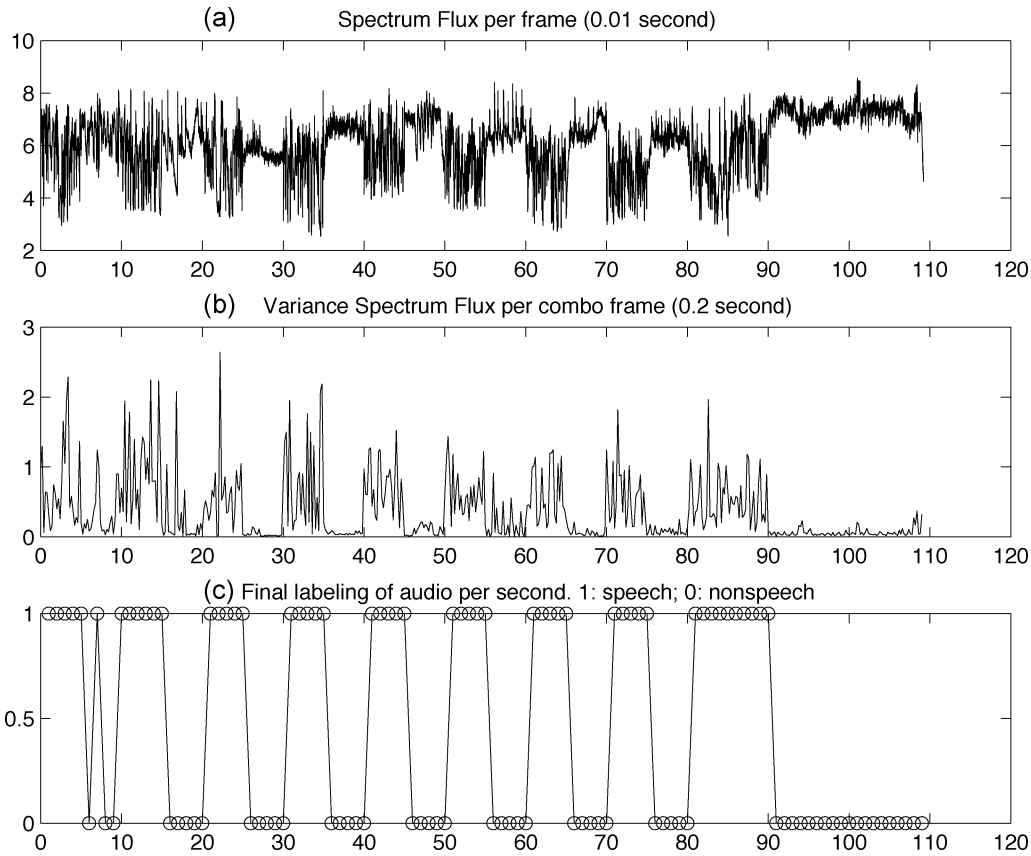


Fig. 2. Speech/nonspeech labeling based on VSF of the 109-s audio file. (a) SF value per frame (0.01 s). (b) VSF value per sub-block (20 frames, 0.2 s). (c) Final labeling of speech/nonspeech per second. 1: speech; 0: nonspeech.

that the variance of the ZCR of speech should be larger than that for music and environmental sounds. Therefore, the same decision process can be applied as that used for VSF by replacing the steps in the dashed box in Fig. 1 with the ZCR calculation.

C. Classification in a Weighted GMM Network

Next, the new classification algorithm is proposed that combines the feature-based methods and model-based methods in an efficient way. Here, the extended-time features VSF and VZCR proposed in the previous section are used. The GMM network (GN) is selected, since it is an efficient model-based classification method that has been considered by others [10], [12]. For the system formulation, three sets of GMM networks are trained: speech and nonspeech; female and male; female broadband, female narrowband, male broadband, male narrowband; and nonspeech. The training data is the DARPA Hub4 BN 1996 training set. For the speech model, the training data is taken from the seven BN focus conditions [30]: F0, F1, F2, F3, F4, F5, FX. For the nonspeech model, the training data is selected from the gaps between the BN speech segments. The training data for narrowband models is from the F2 condition, and data from the other six focus conditions is used for broadband models. A 39-dimensional MFCC feature set is used for the models. The GMMs have between 96 and 256 mixture components depending on training data size and contain diagonal covariance matrices. There are no restrictions on the transition between GMMs/states, so all GMMs are connected to each other, forming a *GMM Network*. The GMM transition probabilities are tuned on the development data as in Section V-A1. The Viterbi algorithm is then implemented to obtain the best classification results. The Viterbi search can be done at the frame level (e.g., 0.01 s per frame; the MFCCs are computed at the frame level) or at a bigger block. We test blocks which are from 0.01 s to 2 s long in duration and find that the 0.75-s to 1.5-s blocks achieve the best performance. In our algorithm, therefore, the Viterbi search is based on 1-s processing blocks in order to match the block size in the classification algorithm based on extended-time features (i.e., VSF and VZCR).

To train the diagonal covariance GMMs, short-time features such as MFCCs must be employed. Although extended-time features cannot be used in model training directly, it is possible to integrate them as a weighting process of the output probabilities of the GMMs. Extended-time features can classify audio into speech versus nonspeech as previously discussed. The weights of the output probabilities of the GMMs are set as follows. If the current segment is classified as speech, then the output probability of the speech GMM for that segment is multiplied by a weight larger than one, and unchanged for the nonspeech model probability. If the segment is classified as nonspeech, then that output probability of the nonspeech model is multiplied by a weight larger than one, and unchanged for the speech model probability. The weights are also tuned on the development data as in Section V-A1. This algorithm can improve classification accuracy significantly, since it combines the strengths of both extended-time features and models in an effective way.

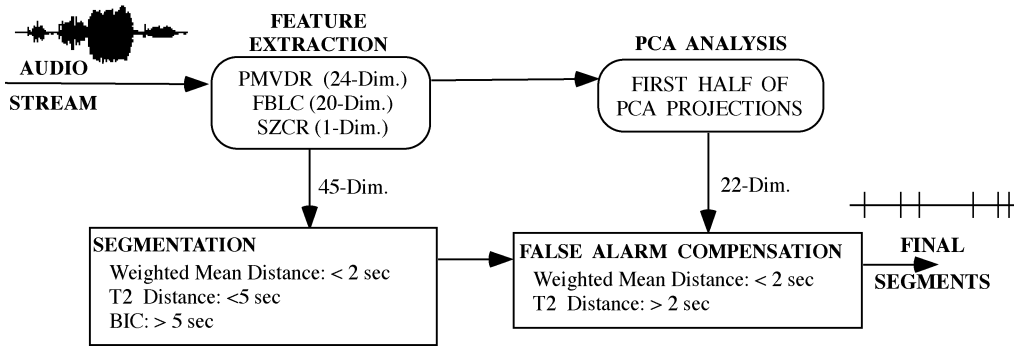


Fig. 3. Block diagram of CompSeg algorithm.

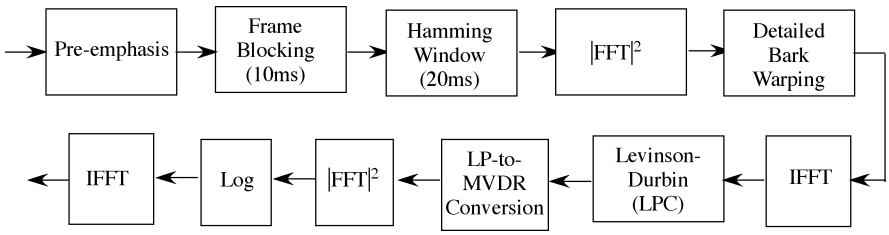


Fig. 4. PMVDR feature extraction process.

IV. CompSeg Algorithm

In this section, a new algorithm for audio segmentation, CompSeg, is proposed. The block diagram of CompSeg algorithm is shown in Fig. 3. The CompSeg includes the following three aspects advances:

1) features specifically designed for segmentation;
2) three distance metrics employed in the analysis window according to its size;
3) specific features and distance metrics for false alarm compensation.

A. Three Features

Here, three new features are considered. They are compared to traditional MFCCs in subsequent evaluations. All features use a 20-ms analysis window with a 10-ms skip frame rate between windows (i.e., 100 frames per second).

1) PMVDR: High-order minimum variance distortionless response (MVDR) models provide better upper envelope representations of the short-term speech spectrum than MFCCs [7]. Furthermore, it has been shown that the MVDR spectrum can be simply obtained from a noniterative computation of the linear prediction (LP) coefficients [7]. A perceptual-based MVDR (PMVDR) feature formulation was developed in [37] and shown to outperform MFCCs for ASR applications. The block diagram of the PMVDR feature extraction process is in Fig. 4. An important trait of PMVDR is that it does not require an explicit filterbank analysis of the speech signal. For the application of speaker segmentation, the order of the LP model is increased to reflect more speaker dependent information in the features. A detailed Bark frequency warping is also applied for better results.

2) SZCR: In Section III-B, the variance of the ZCR was proposed for audio classification. A high ZCR ratio (HZCRR) has also been proposed for audio classification [22]. In experiments, a smoothed ZCR (SZCR) was found to be more efficient for speaker segmentation, which is computed as follows. 1) Compute five sets of ZCR evenly spaced across the analysis window (i.e., one frame) with no intermediate overlap. 2) Next, use the mean of the five sets as the feature of this frame, which reduces the feature variance and thereby increases class separability [7].

3) FBLC: Although in [37], it was suggested that direct warping of the fast Fourier transform (FFT) power spectrum without filterbank processing can preserve almost all the information in the short-term speech spectrum, we find that filterbank processing is more sensitive than other features in detecting speaker changes (i.e., the mismatch between the experimental break points and the actual break points is very small). As such, the FBLC are simply the 20 Mel frequency filterbank log energies coefficients.

From the experiments, the best feature set is the combined feature, i.e., 24-dimensional PMVDR (static, delta, energy), 20-dimensional FBLC, and 1-dimensional SZCR, as shown in Fig. 3.

B. Model Selection Based on the Size of Analysis Window

If audio segments are more than 5 s long, BIC and other distance metric-based methods perform segmentation well [5], [16], [39]. However, in real audio data from BN or two-way conversations, many segments are very short (i.e., less than 5 s). Since BIC and most distance metric-based methods need the second-order statistics (i.e., the covariance), they often suffer in estimation error due to insufficient data.

The Kullback–Leibler distance (KL2) is a popular distance metric in speaker segmentation [29]. If two audio segments can be modeled by multivariate Gaussian distributions $N(\mu_1, \Sigma_1)$

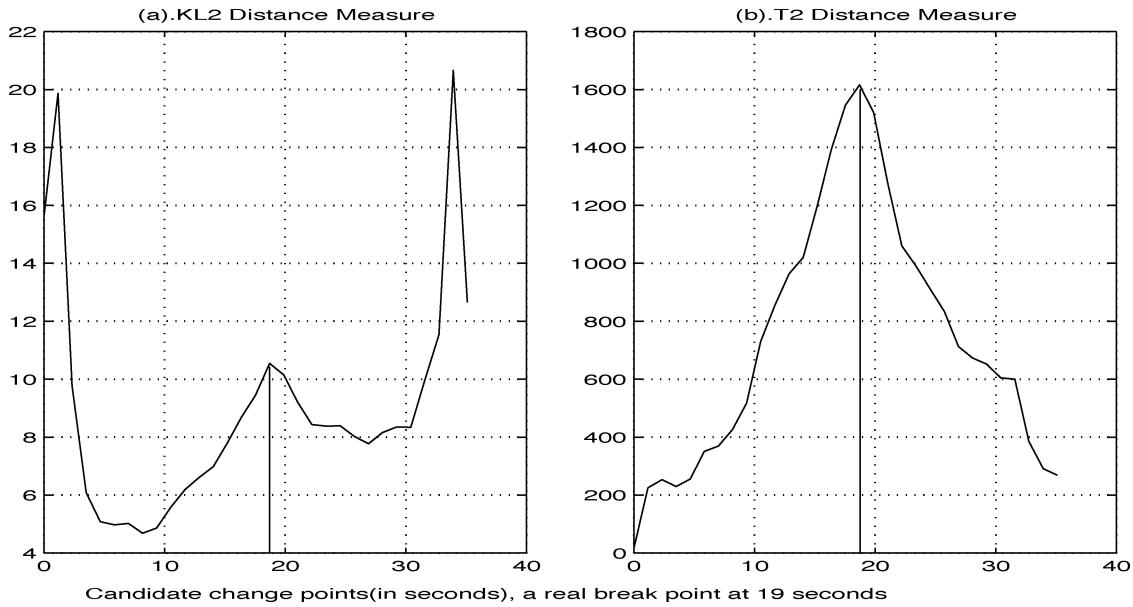


Fig. 5. KL2 and T^2 distance of an audio window, which has a real break point at 19 s.

and $N(\mu_2, \Sigma_2)$, then the KL2 distance between the segments is[2]

$$\mathrm{KL2}_{1,2} = \frac{1}{2}(\mu_1 - \mu_2)^T \left(\Sigma_1^{-1} + \Sigma_2^{-1}\right)(\mu_1 - \mu_2) + \frac{1}{2} tr\left(\Sigma_1^{-1}\Sigma_2 + \Sigma_2^{-1}\Sigma_1 - 2I\right). \quad (6)$$

Fig. 5(a) shows the KL2 distance of a 35-s audio stream which has only one real break point at 19 s. Here, the KL2 distance measure of the first and final 5 s are not correct. This occurs because of insufficient data in the estimation of the covariance when the segment is shorter than 5 s. In contrast, Fig. 5(b) shows that the T^2 distance measure detects the break point accurately with no initial or trailing edge effects.

The idea of using the Hotelling T^2-statistic [3], [4], [34], [39] for speaker segmentation is that for two audio segments, if they can be modeled by multivariate Gaussian distributions: $N(\mu_1, \Sigma_1)$ and $N(\mu_2, \Sigma_2)$, we assume their covariances are equal but unknown, then the only difference between them is the mean values reflected in the T^2 distance as

$$T^2 = \frac{ab}{a+b}(\mu_1 - \mu_2)^T \Sigma^{-1}(\mu_1 - \mu_2) \quad (7)$$

where a and b are the number of frames within each of the audio segments, respectively. Under the equal covariance assumption, more data can be used to estimate the covariance and reduce the impact of insufficient data in the estimation. This is the primary reason why the T^2 distance measure can detect the break point accurately in Fig. 5(b). If the processing audio window is shorter than 2 s, even a global covariance will suffer from insufficient estimation. We can, therefore, further assume the global covariance to be an identity matrix, in which case this is termed as weighted mean distance. Weighted mean distance is applied for processing audio windows shorter than 2 s and T^2 distance for those shorter than 5 s and longer than 2 s. This distance measure is termed as T^2-mean. Fig. 6 clearly shows that if there is a break point in the processing window, the distance measure has *one and only one prominent peak*. Therefore, the T^2-mean can be used to detect the break point in a short processing window (<5 s) effectively. As the window grows in duration (>5 s), the covariance can be estimated more accurately and BIC is better than T^2 in speaker turn detection. Therefore, BIC is applied to detect break points directly as in [5], [39]. The distance metric scheme in the segmentation stage of CompSeg is as follows:

1) $L_w < 2$ s: weighted mean distance;
2) $2\ \mathrm{s} \le L_w < 5$ s: T^2 distance;
3) $L_w \ge 5$ s: traditional BIC.

where the L_w is the length of the processing audio window, T^2-mean is 1 and 2 together.

C. False Alarm Compensation

1) Audio Clustering Based False Alarm Compensation: It is common that speech from the same speaker might appear multiple times in an audio stream. In general, it would be useful to pool the homogeneous data from the same speaker for subsequent processing (e.g., speaker adaptation, speaker identification, etc.). The application of traditional BIC and distance measures for a hierarchical clustering is straightforward [5], [9], [12], [15]. The clustering is implemented in a bottom-up framework, where each segment is a node, and the distance is calculated between all nodes and apply BIC to examine node pairs with the nearest distance if they can be merged. If they are homogeneous, they are merged to a single new node and the distance matrix is recalculated. If they are not homogeneous, then consider the next nearest node pair. This procedure continues until all nodes have been examined.

2) Transformed Combined Feature and New Distance Metric for False Alarm Compensation: If the two mergable nodes are adjacent segments in the audio clustering routine, it means that

[2]The basic assumption for BIC, other distance metric approaches, and our method is that there is at most one real break point in the processing window. This assumption is reasonable for speaker segmentation.

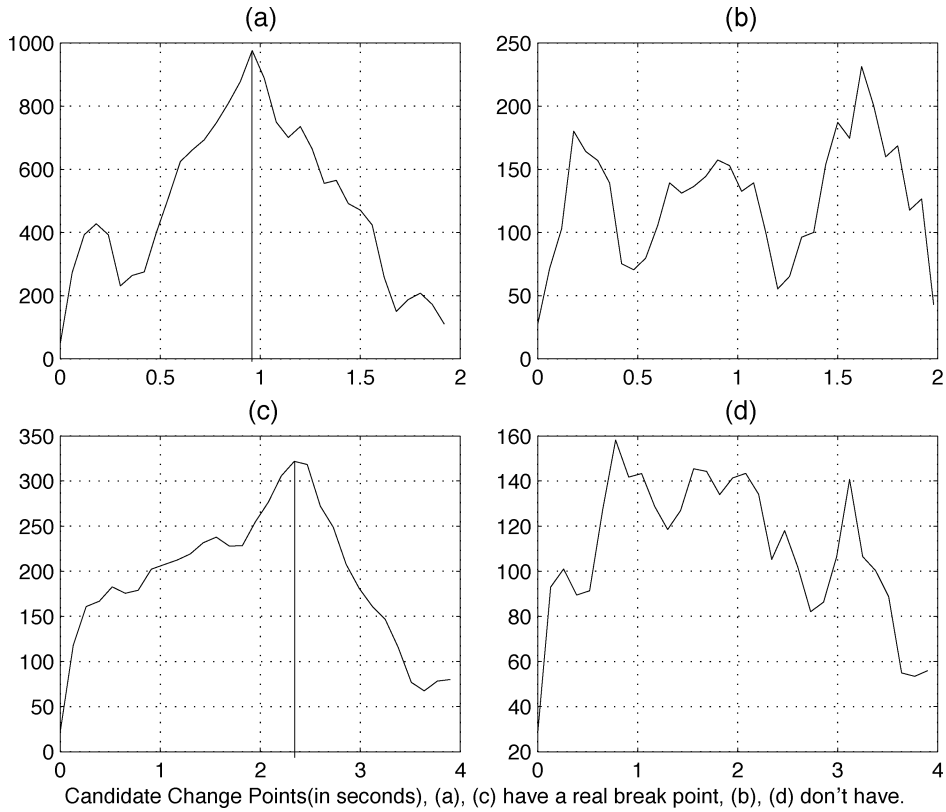


Fig. 6. T^2-mean distance of processing audio windows, x-axis is the window length (in second), y-axis is the distance. (a), (b) is the weighted mean distance; (c), (d) is the T^2 distance. (a), (c) have one real break point at 0.9 second, 2.4 second respectively, (b),(d) have no true break points.

the false alarm rate can be compensated. However, a false alarm compensation method based on existed clustering techniques is not powerful, because it cannot compensate the false alarm caused by short segments due to reduced data size. Conceptually, false alarm compensation is similar to classification. Here, the distance between two adjacent segments is calculated, and if the distance is below a threshold, then they belong to the same class, (i.e., a false alarm break point is found), otherwise they are from different classes. A weighted mean distance metric is applied for short segments and regular T^2 with covariance matrix estimation for long segments. This scheme can compensate the false alarm rate significantly with little cost to the miss rate. The distance metric in the false alarm compensation stage of CompSeg is as follows:

1) $L_s < 2$ s: weighted mean distance;
2) $L_s \geq 2$ s: T^2 distance.

where L_s is the total length of the two adjacent segments.

The feature used in false alarm compensation can be different than that used in segmentation. Here, the first half of the principal components analysis (PCA [17]) projections of the combined feature: PMVDR + FBLC + SZCR (PMVDR: 24 dimensions, FBLC: 20 dimensions, SZCR: 1 dimension) is used. The PCA projection has more discriminative power than the original feature since it determines and rank orders those dimensions that occupy a larger percentage of the data variability, making it suitable in this classification-like task. Integrating all the advances proposed in Section IV together, the CompSeg algorithm is implemented as in Fig. 3.

TABLE I
CLASSIFICATION ACCURACY USING THE CONSTRUCTED 515-s AUDIO FILE

Scheme	Frame Accuracy	
	Speech	Non-speech
VSF	93.3%	86.3%
VZCR	92.0%	90.9%
GN	97.3%	92.1%
VSF-WGN	100%	93.7%
VZCR-WGN	99.3%	94.8%

V. EXPERIMENTS

For the experiments, the evaluation data is drawn from TIMIT [32], CU-MOVE [13], DARPA Hub4 BN 1996 training data and 1997 evaluation data [30], and NGSW data [25]. This test data includes the following:

1) different structures of audio: interviews, reports, debates, etc.;
2) various recording equipment: microphone, telephone, Edison cylinder disks, TV/radio;
3) various background noise: music, audience laughing, clapping, automobile noise (road, wind, turn signals), etc.

The experiments are composed of four parts:

1) classification evaluation;
2) feature evaluation in speaker segmentation;
3) CompSeg evaluation;
4) WGN and CompSeg evaluation with NGSW data.

All the parameters for audio classification and segmentation (i.e., the threshold for VSF and VZCR, the weights for WGN, the transition probabilities for GN, the threshold for BIC, etc.)

TABLE II
SPEECH/NONSPEECH CLASSIFICATION IN THE WGN

Scheme	Precision		Recall		Total Accuracy
	Speech	Non-Speech	Speech	Non-Speech	
GN	98.08%	28.73%	94.86%	52.81%	93.4%
VSF	98.74%	26.10%	92.21%	70.09%	91.5%
VSF-WGN	98.22%	56.64%	98.36%	54.63%	96.7%
VZCR	98.80%	15.09%	83.63%	74.09%	83.5%
VZCR-WGN	97.90%	35.83%	96.68%	47.26%	94.9%
VSF+VZCR-WGN	**98.22%**	**58.42%**	**98.48%**	**54.44%**	**96.9%**

are tuned on the development set as stated following and left unchanged for the standard data set evaluation.

A. Classification Evaluation

1) Parameter Tuning in the Development Set: In order to set up the parameters for real data evaluation, an audio file is artificially generated, which is designed to include as many acoustic events as possible. A set of five broadband female speech clips, five female speech clips with background music, and five narrowband (i.e., telephone) female speech clips are selected. The process is also repeated with male speech clips, where each clip is 5 s in duration, resulting in mixed speech audio stream of duration 150 s. The speech clips are drawn from DARPA Hub4 96 BN training corpus and across all seven focus conditions (F0, F1, . . ., FX [30]). Next, 36 music clips, 36 environmental sound clips, and one silence clip are selected, where each clip is also 5 s in duration. The music clips include a variety of music types such as jazz, rock, blues, pop, classical, etc.. The environmental sounds also include a variety of events such as audience clapping and shouting, coughing, laughing, automobile noise (e.g., engine, beeps, noise from car with windows open), electronic noise, and others. The sequence of music clips and environmental sound clips are concatenated after the mixed speech clip sequence. The total length of the resulting audio file is 515 s.

Table I shows the classification accuracy of VSF, VZCR, basic GN, VSF weighted GN, and VZCR weighted GN in this development data set. From Table I, it is observed that the VSF can achieve better performance in the speech part than the VZCR; however, it is worse than the VZCR in the nonspeech part. It is also observed that the classification power of extended-time feature (e.g, VSF, VZCR) and the GN is additive.

2) Evaluation on the Hub4 Data: The DARPA Hub4 BN 1997 corpus evaluation data is used in the classification evaluation. This data set of three hours includes many audio types: broadband (microphone) speech, narrowband (telephone) speech, speech on music, various environment sounds, and multispeaker change, and is, therefore, suitable for classification evaluation. The GN algorithm will represent the baseline. The precision and recall of class C are defined as

$$\text{Precision} = \frac{\#\text{ of frames correctly labeled as } C}{\#\text{ of frames labeled as } C} \quad (8)$$

$$\text{Recall} = \frac{\#\text{ of frames correctly labeled as } C}{\#\text{ of frames in } C}. \quad (9)$$

Table II demonstrates that although the classification performance of extended-time features (i.e., VSF and VZCR) is less accurate than that of GN, they can still contribute to improving GN classification. The VSF weighted GN ("VSF-WGN" in Table II) can outperform both GN and VSF algorithms; The VZCR weighted GN ("VZCR-WGN" in Table II) can also outperform both GN and VZCR algorithms. Finally, the combined VSF and VZCR weighted GN ("VSF + VZCR – WGN" in Table II) outperforms the baseline GN at **ALL** levels (precision and recall in both speech and nonspeech parts) and it improves the total frame accuracy from 93.4% to 96.9%. The relative improvement in error reduction is over 50%. Table II also shows that if the VSF and VZCR are combined together in the weighted GN, it only achieves slightly better results than the single VSF weighted GN classifier. This suggests that the contributions from VSF and VZCR for classification do overlap. Since major part of the evaluation data is speech and VSF can achieve better performance versus VZCR, it partially confirms the argument in Section V-A1 that VZCR has better classification power versus VSF for nonspeech segments, but is less successful than VSF in speech classification.

TABLE III
AUDIO CLASSIFICATION WITHOUT/WITH SEGMENTATION

Discrimination Type	Frame Accuracy	
	Without Seg	With Seg
Female/Male	86.7%	92.9%
Female-Broadband/Male-Broadband/Female-Narrowband/Male-Narrowband/Non-Speech	81.0%	82.1%

Since speech blocks from segmentation are much longer and more homogeneous (same speaker in a consistent acoustic environment) than predefined processing windows (1-s duration is used for classification without segmentation) for classification, it is reasonable to expect the classification result based on segmentation to be better than classification without segmentation since the presegmentation can provide the processing windows (i.e., segments from the segmentation) for classification. That is to say, with presegmentation, we classify the whole segments into classes regardless of the length of the segments. Table III demonstrates such improvement by applying segmentation, where the test set is also Hub4 BN 97 evaluation data and the classification is implemented using GMM networks.

Since there are nonspeech parts, the VSF or VZCR weighted GN can still be applied in the five-state classification framework (female-broadband/ male-broadband/ female-narrowband/ male-narrowband/ nonspeech). From Table IV, the VSF or VZCR weighted GN algorithm outperforms the classification based on segmentation algorithm.

It would be reasonable to question why classification with the WGN should be better than that based on segmentation. It

TABLE IV
FIVE-STATE CLASSIFICATION IN A WGN

Scheme	Frame Accuracy
GMM Network(GN)	81.0%
Based on Segmentation	82.1%
VSF-WGN	83.4%
VZCR-WGN	82.6%
VSF+VZCR-WGN	**83.6%**

is suggested that in the segmentation phase, missing the actual break points has a pronounced impact on classification, since it causes nonhomogeneous audio segments to be combined together, corrupting the statistical analysis and, therefore, classification errors will occur. Furthermore, GMM networks can make decisions based on short duration audio blocks, so the long homogeneous segments from segmentation will not provide much benefit. However, this additional data can still improve classification. Restated, if the nonspeech parts in the audio stream need to be classified, the WGN (i.e., VSF or VZCR weighted) can be applied; otherwise, classification based on segmentation is an acceptable method.

B. Feature Evaluation in Audio Segmentation

Next, the selection of speech features for segmentation is considered. The pure BIC segmentation algorithm is employed here. The goal is to determine which feature is most suitable for speaker segmentation while maintaining successful performance for a range of background noise conditions. To determine this, the following experiments are designed. First a set of simple audio files are constructed where each is composed of noise, speech, noise, speech,. . ., noise segments sequentially. For each stream, there is a total of ten speech sentences and 20 real break points. The speech sentences are drawn from the same speaker from the TIMIT corpus and each is roughly 3 s in duration. The noise inserted between each sentence is 3 s in duration as well. A set of five female and five male audio files are constructed for each noise type. A total of 14 noise types are considered, which come from the CU-move in-vehicle database [13] and an earlier speech recognition study [14]. The evaluation considers 16 features: MFCC family [FFT-based, LP-based, with/without variable cepstrum mean normalization (VCMN) [8], 13/26 dimensions, RCC (root cepstrum coefficients)], TEO-CB-AutoEnv (Teager energy operator, critical band, autocorrelation envelope) [19], [41], LPC, LSP, pitch-energy, perceptual MVDR cepstrum coefficients (PMCC) [36] and the three new features from Section IV-A: PMVDR, SZCR, and FBLC, where SZCR is one-dimensional, so it is only employed in the feature combination. Table V shows the top three performing features in the 14 noise types.[3] Fig. 7 shows time versus frequency response of four typical noise types. Table VI shows the average performance of the features across all the noise conditions. Table V shows that PMVDR and FBLC normally occupy the top three positions of the features considered. The average mismatch is computed on the correctly labeled break points and defined as

$$\overline{L_{Mis}} = \frac{\sum_{i=1}^{N} L_{Mis}(C_i, A_i)}{N} \tag{10}$$

[3]See Appendix A for a description of the noise types.

where $\overline{L_{Mis}}$ is the average mismatch, $L_{Mis}(C_i,\ A_i)$ is the mismatch between the correctly labeled break point C_i and the actual break point A_i, N is the total number of the correctly labeled break points. If $L_{Mis}(C_i,\ A_i) < 1$ s, C_i is considered to be the correctly labeled break point.

From Table VI, PMVDR and FBLC are better than other features in the sense of noise robustness. It would be useful to ask why FBLC features are more sensitive to speaker change than MFCCs. We believe that the FBLC retains the correlation across the filter outputs, which are useful for speaker segmentation. However, retaining that feature correlation is harmful for speech recognition, so the discrete cosine transform (DCT) is applied to the FBLC feature to decorrelate and obtain the well-known MFCCs.

Further evaluation is performed using the DARPA Hub4 BN data. Table VII shows that the PMVDR feature outperforms MFCCs at all levels, and FBLCs result in a small average mismatch which implies they are sensitive to changes between speakers and environments. The experiment demonstrates that PMVDRs are better than MFCCs in speaker segmentation and FBLCs are more sensitive than MFCCs in speaker change detection. Because the PMVDRs do not apply filterbank processing, combining PMVDRs and FBLCs could improve performance. Also, the SZCR encodes information directly from the waveform which is incorporated into the combination as well. Finally, the combined feature set consists of the 24 features from PMVDR (the last two delta PMVDR coefficients have no impact on the segmentation performance in the experiments, so they are dropped in the feature combination), all 20 features from FBLC, and one SZCR (i.e., a 45-dimensional set). The features are normalized to zero mean and unit variance for improved discrimination ability [7]. Table VII shows the advantages of employing feature combination. Other prosodic features such as pitch were also considered, but the results showed little improvement. The reason may be because pitch only encodes information from voiced speech, and contains no information from unvoiced speech and noise making it less effective for segmentation.

C. CompSeg Evaluation

Next, the proposed segmentation algorithm, CompSeg, is evaluated for each processing step on the DARPA Hub4 BN 96/97 evaluation data in order to identify performance both individually and for the overall system. The parameters for the CompSeg are tuned from 2-h data drawn from the Hub4 BN 96 training data set.

1) T^2-Mean and False Alarm Compensation Evaluation: The evaluation data used in this section is drawn from the DARPA Hub4 BN 1996 evaluation data. The performance of the new segmentation scheme T^2-mean is shown in Table VIII, where a 24-dimensional PMVDR feature set is used in both baseline and T^2-mean segmentation. The baseline system in this experiment uses BIC only. With this advance, the false alarm rate is significantly reduced, and there is a 2.2% absolute improvement in the miss rate, with 2.0% coming from the short segments. This suggests that the contribution of T^2-mean is mainly derived from short duration turn detection.

In order to apply the proposed false alarm compensation routine, an initial segmentation is applied to find all possible

TABLE V
TOP THREE PERFORMING FEATURES IN NOISE ENVIRONMENTS. FA: FALSE ALARM RATE; MISS: MISS RATE; MMATCH: AVERAGE MISMATCH (ms)

Noise	Top 1			Top 2			Top 3		
	FA	Miss	MMatch	FA	Miss	MMatch	FA	Miss	MMatch
AIR	RCC			**FBLC**			**PMVDR**		
	1.0%	0.0%	126.78	0.0%	0.0%	128.81	0.0%	0.0%	130.68
AWG	RCC			**FBLC**			**PMVDR**		
	0.0%	0.0%	120.58	0.0%	0.0%	121.59	0.0%	0.0%	123.95
CRA	RCC			**FBLC**			**PMVDR**		
	0.0%	0.0%	112.44	0.0%	0.0%	112.54	0.0%	0.0%	119.78
FLN	RCC			**FBLC**			**PMVDR**		
	0.0%	0.0%	108.75	0.0%	0.0%	123.77	0.0%	0.0%	125.56
HEL	**FBLC**			MFCC26			**PMVDR**		
	0.0%	0.0%	122.72	0.0%	0.0%	130.31	0.0%	0.0%	131.27
HWY	**FBLC**			**PMVDR**			PMCC		
	0.0%	0.0%	120.65	0.0%	0.0%	123.98	0.0%	0.0%	129.00
LCI	**PMVDR**			RCC			**FBLC**		
	0.0%	0.0%	122.62	0.0%	0.0%	123.51	0.0%	0.0%	124.56
LCR	**FBLC**			**PMVDR**			PMCC		
	0.0%	0.0%	111.16	0.0%	0.0%	113.14	0.0%	1.0%	113.62
PS2	**FBLC**			MFCC26-VCMN			RCC		
	0.5%	0.0%	131.57	0.5%	0.0%	140.16	0.0%	0.0%	142.38
SUN	MFCC26-VCMN			**FBLC**			**PMVDR**		
	0.0%	0.0%	125.74	0.0%	0.0%	131.44	0.0%	0.0%	131.85
Blazer AC	**FBLC**			**PMVDR**			RCC		
	0.0%	0.0%	130.14	0.0%	0.0%	132.60	0.0%	0.0%	137.74
Blazer Truck	**FBLC**			**PMVDR**			MFCC26		
	0.0%	0.0%	113.77	0.0%	0.0%	116.94	0.0%	0.0%	118.58
Blazer Turn	**PMVDR**			**FBLC**			PMCC		
	0.0%	0.0%	120.16	0.0%	0.0%	120.13	0.0%	0.0%	127.93
Blazer Wind	RCC			**PMVDR**			**FBLC**		
	1.0%	0.0%	110.04	0.0%	0.0%	113.46	0.0%	0.0%	117.11

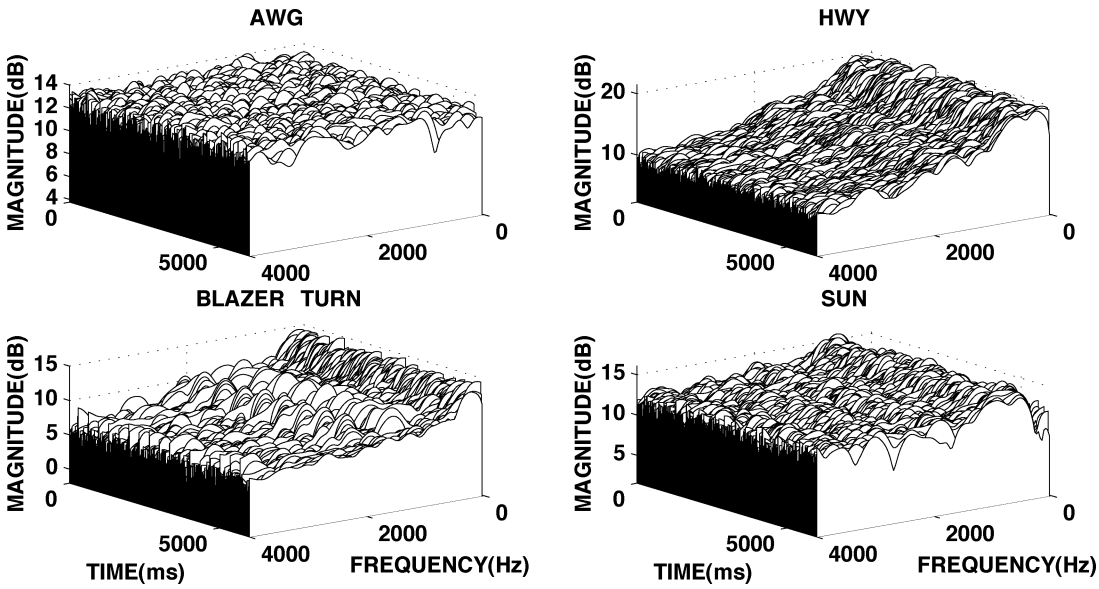


Fig. 7. Time versus frequency contours of four typical noise types.

break points regardless of the false alarm rate. Table IX shows that the false alarm compensation scheme is very effective. In the segmentation stage, the feature is a 24-dimensional MFCC set [24]. In the false alarm compensation stage, the feature is the first half of the PCA projections of the combined feature: PMVDR + FBLC + SZCR (i.e., a 22-dimensional feature set, Section IV-C2). The baseline system uses a 24-dimensional MFCC set with traditional BIC.

2) DARPA Hub4 Standard Evaluation: The DARPA Hub4 BN 1997 evaluation data was used for performance assessment. The set contains 3 h of broadcast news data, with 584 break points that includes 178 short segments (<5 s). The proposed CompSeg algorithm uses the PMVDR, SZCR, FBLC combined features (i.e., a 45-dimensional set as in Table VII), applies T^2-mean distance measure for segments of duration shorter than 5 s, BIC model selection for longer duration segments,

TABLE VI
AVERAGE PERFORMANCE OF ALL THE FEATURES. [(∗) MEANS DIMENSIONS OF THE FEATURE]

Feature	MFCC(13)	MFCC-VCMN(13)	LPMFC(13)	LPMFC-VCMN(13)	MFCC(26)	MFCC-VCMN(26)	LPMFC(26)	LPMFC-VCMN(26)
FA	26.6%	9.5%	39.3%	24.9%	0.1%	0.5%	5.1%	3.0%
Miss	0.0%	0.0%	0.0%	0.1%	0.0%	1.6%	0.0%	0.1%
MMatch	357.96	229.95	411.36	324.59	134.24	135.48	194.24	177.25
Feature	LPC(10)	LSP(10)	Pitch-Energy(2)	TEO-CB-AutoEnv(17)	PMCC(26)	RCC(26)	**PMVDR(26)**	**FBLC(20)**
FA	27.0%	7.0%	36.5%	11.8%	0.0%	0.3%	**0.0%**	**0.0%**
Miss	0.0%	0.4%	0.0%	1.0%	0.1%	0.0%	**0.0%**	**0.0%**
MMatch	122.15	334.46	210.04	438.06	240.42	133.40	**127.02**	**124.93**

TABLE VII
FEATURE EVALUATION. "()" IS THE RELATIVE IMPROVEMENT, FA: FALSE ALARM RATE (%); MIS: MISS DETECTION RATE (%); MMATCH: AVERAGE MISMATCH (ms). (NOTE: KEY IS THE SAME AS THE FOLLOWING TABLES)

Feature	FA	MIS	MMatch
MFCC(26-D)	29.6%	25.0%	298.47
FBLC(20-D)	29.8% (-0.7%)	25.3% (-1.2%)	266.80 (10.6%)
PMVDR(26-D)	25.9% (12.5%)	24.9% (0.4%)	284.29 (4.8%)
Combine 45-D	23.8% (19.6%)	24.3% (2.8%)	265.06 (11.2%)

TABLE VIII
EVALUATION OF T^2-MEAN SEGMENTATION

Scheme	FA	MIS	MMatch
Baseline	27.6%	27.4%	277.50
T^2-Mean	23.5% (14.9%)	25.2% (8.0%)	281.21 (-1.3%)

TABLE IX
EVALUATION OF FALSE ALARM COMPENSATION SCHEME

Scheme	FA	MIS	MMatch
Baseline	44.2%	18.7%	307.83
FA-COMP	23.8% (23.5%)	21.3% (-13.9%)	292.28 (5.1%)

TABLE X
EVALUATION OF CompSeg WITH Hub4 BN 1997 EVALUATION DATA

Algorithm	FA	MIS	MMatch
Baseline	26.7%	26.9%	293.02
CompSeg	21.1% **(21.0%)**	20.6% **(23.4%)**	262.99 **(10.2%)**

and finally applies the false alarm compensation postprocessing routine. The block diagram of CompSeg algorithm is shown in Fig. 3. The improvement using these advances is shown in Table X. For **ALL** metrics, performance improves significantly on the Hub4 data. The baseline system uses 24-MFCCs and traditional BIC only.

D. NGSW Data Evaluation

Our final goal of audio stream parsing is to determine automatic and effective audio classification and speaker segmentation for a spoken document retrieval system (i.e., the SpeechFind system [31] for NGSW), which focuses on web-based access to the largest collection of historical audio materials in the U.S. (e.g., as much as 60 000 hours of materials from the past 100 years). Therefore, the final evaluation of the WGN and CompSeg employs NGSW data.

TABLE XI
NGSW DATA CLASSIFICATION IN WGN

Scheme	**Frame Accuracy**	
	7-decades	1960s
GMM Network(GN)	73.3%	96.0%
VZCR only	79.6%	82.2%
VZCR-WGN	**86.5%**	**97.8%**

Two sets of audio materials from the NGSW corpus are selected. The first consists of audio samples from seven decades (1940s, 1950s, 1960s, 1970s, 1980s, 1990s, 2000s), where each clip is a typical audio representation of topic, recording media/equipment and speaker content for that period.[4] The second set is an audio stream from the 1960s which is more topic specific.

Here, VZCR is used as the extended-time feature for GMM network weighting. Table XI clearly shows the effectiveness of the WGN algorithm. Most clips in the seven-decades data were recorded outdoors, thus containing diverse and varying levels of noise. Some speech was misclassified as noise, and some audience noise was misclassified as speech, though overall frame accuracy was 86.5%. The 1960s data consists of indoor broadcast news, which is less noisy than the seven-decades data. Since the GMMs were trained with the DARPA Hub4 BN 1996 training data, the 1960s test data should match the models well. However, the seven-decades data does not match the system models at all in terms of topic, speaker, and recording environment/equipment. Such differences can be reflected from the frame accuracy of GN classification of 96.0% versus 73.3%. However, the extended-time feature shows the consistency in classification with an 82.2% versus 79.6% performance rate for clean or noisy conditions. From this observation, we conclude that GN classification performance is more sensitive to the data. If the test data is quite different from the training data, some form of model adaptation [e.g., maximum likelihood linear regression (MLLR) [20]] should be applied to the GMMs. Second,

[4]The content consists of: 1940s: Chicago Roundtable on West European reconstrucion after World War II; 1950s: President Hoover's speech on the relationship between the U.S. and Russia; 1960s: Report of the Cuban Missile Crisis; 1970s: President Nixon's speech on the Watergate Tapes; 1980s: President Reagan's speech on his landslide reelection in 1984; 1990s: President George Bush's speech on the International Drug Control Cartel in Columbia; 2000s: News conferences regarding the hand counting of votes for the U.S. President election in Florida.

TABLE XII
NGSW DATA SEGMENTATION IN CompSeg

Rate	7-decades		1960s	
	Baseline	CompSeg	Baseline	CompSeg
FA	18.2%	0.0%	27.8%	5.6%
Miss	0.0%	0.0%	11.1%	0.0%
MMatch(ms)	149	132	166	117

the extended-time features are robust to the acoustic environments; however, their classification power is limited, and that by combining their effects with GN (i.e., WGN) results in an effective overall classification method.

Table XII compares the segmentation evaluation between the seven-decades and 1960s NGSW data. The baseline uses 24-MFCCs plus traditional BIC segmentation. Since the acoustic changes in the seven-decades and 1960s data are large, it is not hard to detect the break points in the audio streams. However, the abrupt acoustic changes in the audio will also cause false alarms, which is the reason for reduced miss rates and much higher false alarm rates. Table XII shows that the CompSeg solution outperforms the baseline system at all three levels.

VI. CONCLUSION

This study has considered advances in unsupervised audio classification for LVCSR and speaker segmentation for multi-speaker change detection. Two new extended-time features, VSF and VZCR were proposed for audio classification and a novel classification algorithm: WGN, was presented (Section III). VSF and VZCR were shown to be robust and effective for speech/non-speech classification (Section V-A1). The WGN classification algorithm combines a feature-based method and model-based method in a compact and reliable way. The WGN improves the frame accuracy from 93.4% to 96.9% over traditional GMM networks and outperforms the baseline system at **ALL** levels (Section V-A2). We note that other extended-time features could be considered in the future. Feature selection was shown to be very important for speaker segmentation. A broad range of features (e.g., 16 features) were evaluated in 14 types of adverse noise environments, resulting in a set of noise robust features: PMVDR, FBLC, SZCR, along with their combinations (Section V-B). It was also shown that a systematic set of advances integrated into the new segmentation algorithm, CompSeg, can achieve effective unsupervised speaker segmentation, especially for short duration segments. Significant improvement was achieved in **ALL** metrics over a traditional BIC with MFCC-based segmentation algorithm (Section V-C2).

APPENDIX A

Here, a brief summary of the noise types considered in Section V-B are presented:[5]

AIR aircraft cockpit noise from a Lockheed C130 cargo plane flying at 25 000 ft;
AWG white Gaussian noise;
CRA noise from a large city with rain falling;
FLN flat voice communication channel noise from a U.S. telephone channel;
HEL noise recorded from the ground with a helicopter fly-by taking place;
HWY noise from a car with windows closed traveling on highway at 65 mi/h;
LCI large city noise recorded at street level;
LCR large crowd noise in an auditorium;
PS2 cooling fan noise from an IBM PS2 model 80 computer;
SUN cooling fan noise from a Sun Sparcstation 330 computer;
Blazer AC SUV (Blazer) driving at less than 45 mi/h with air conditioning on and windows closed;
Blazer Truck SUV (Blazer) driving 45 mi/h with windows open 2 in and truck passing outside;
Blazer Turn SUV (Blazer) driving with turn signal on and windows closed;
Blazer Wind SUV (Blazer) driving with windows open 2 in at 65 mi/h on highway;

[5]A number of these noise types were used in [14].

APPENDIX B

Major acronyms in this paper.

BIC Bayesian information criterion.
BN Broadcast news.
CompSeg Compound segmentation.
FBLC Filterbank log energy coefficients.
(W)GN (Weighted) GMM network.
LP Linear prediction.
LSP Line spectral pair.
(P)MVDR (Perceptual) minimum variance distortionless response.
NGSW The National Gallery of the Spoken Word.
PMCC Perceptual mvdr-based cepstral coefficients.
RCC Root cepstrum coefficients.
(V)SF (Variance) of the spectrum flux.
SZCR Smoothed zero-crossing rate.
VCMN Variable cepstrum mean normalization.
(V)ZCR (Variance) of the zero-crossing rate.

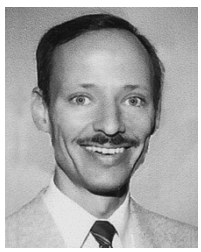
REFERENCES

[1] A. Adami, S. Kajarekar, and H. Hermansky, "A new speaker change detection method for two-speaker segmentation," in *Proc. ICASSP*, vol. 4, Orlando, FL, 2002, pp. 13–17.
[2] J. Ajmera and I. McCowan, "Speech/music discrimination using entropy and dynamism features in a HMM classification framework," IDIAP, Martigny, Switzerland, Tech. Rep. IDIAP-RR 01-26, 2001.
[3] T. Anderson, *An Introduction to Multivariate Statistical Analysis*. New York: Wiley, 1958, pp. 101–125.
[4] M. Cettolo and M. Federico, "Model selection criteria for acoustic segmentation," in *Proc. ISCA ITRW ASR2000 Workshop*, Paris, France, Sep. 2000, pp. 221–227.
[5] S. Chen and P. Gopalakrishnan, "Speaker, environment and channel change detection and clustering via the Bayesian information criterion," in *Proc. Broadcast News Transcr. Under. Workshop*, Lansdowne, VA, 1998, pp. 127–132.
[6] J. R. Deller, J. H. L. Hansen, and J. G. Proakis, *Discrete-time processing of speech signals*, 2nd ed. New York: IEEE Press, 2000, pp. 245–251.
[7] S. Dharanipragada and B. Rao, "MVDR-based feature extraction for robust speech recognition," in *Proc. ICASSP 2001*, vol. 1, Salt Lake City, UT, 2001, pp. 7–11.
[8] M. J. F. Gales and S. J. Young, "Cepstral parameter compensation for HMM recognition in noise," *Speech Commun.*, vol. 12, pp. 231–239, 1993.
[9] J.-L. Gauvain, L. Lamel, and G. Adda, "Partitioning and transcription of broadcast news data," in *Proc. ICSLP*, vol. 2, Sydney, Australia, Dec 1998, pp. 1335–1338.

[10] J.-L. Gauvain, L. Lamel, and G. Adda, "The LIMSI broadcast news transcription system," *Speech Commun.*, vol. 37, pp. 89–108, 2002.
[11] A. H. Gray Jr. and J. D. Markel, "Distance measures for speech processing," *IEEE Trans. Acoust., Speech Signal Process.*, vol. ASSP-24, no. 5, pp. 380–391, Oct. 1976.
[12] T. Hain, S. Johnson, A. Tuerk, P. Woodland, and S. Young, "Segment generation and clustering in the HTK broadcast news transcription system," in *DARPA Broadcast News Transcr. Under. Workshop*, Lansdowne, VA, 1998, pp. 133–137.
[13] J. H. L. Hansen, X. Zhang, M. Akbacak, U. Yapanel, B. Pellom, W. Ward, and P. Angkititrakul, "CU-Move: Advanced in in-vehicle speech systems for route navigation," in *DSP for In-Vehicle and Mobile Systems*, H. Abut, J. H. L. Hansen, and K. Takeda, Eds. New York: Springer-Verlag, 2004.
[14] J. H. L. Hansen and L. Arslan, "Robust feature estimation and objective quality assessment for noisy speech recognition using the credit card corpus," *IEEE Trans. Speech Audio Process.*, vol. 3, no. 3, pp. 169–184, May 1995.
[15] M. Harris, X. Aubert, R. Haeb-Umbach, and P. Beyerlein, "A study of broadcast news audio stream segmentation and segment clustering," *EuroSpeech*, vol. 3, pp. 1027–1030, Sep 1999.
[16] S. Johnson, "Speaker tracking," M.S. thesis, Eng. Dept., Cambridge Univ., U.K., 1997.
[17] I. T. Jolliffe, *Principal Component Analysis*. New York: Springer-Verlag, 1986.
[18] B. Kedem, "Spectral analysis and discrimination by zero-crossings," *Proc. IEEE*, vol. 74, no. 11, pp. 1477–1493, Nov. 1986.
[19] J. F. Kaiser, "Some useful properties of teager's energy operator," in *Proc. ICASSP*, vol. 3, Minneapolis, MN, Apr. 1993, pp. 149–152.
[20] C. J. Leggetter and P. C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," *Comput. Speech Lang.*, vol. 9, pp. 171–185, 1995.
[21] S. Z. Li, "Content-based audio classification and retrieval using the nearest feature line method," *IEEE Trans. Speech Audio Process.*, vol. 8, no. 5, pp. 619–625, Sep. 2000.
[22] L. Lu, H. Zhang, and H. Jiang, "Content analysis for audio classification and segmentation," *IEEE Trans. Speech Audio Process.*, vol. 10, no. 7, pp. 504–516, Oct. 2002.
[23] L. Lu and H. Zhang, "Speaker change detection and tracking in real-time news broadcasting analysis," in *Proc. ACM Multimedia*, Juan-les-Pins, France, Dec. 2002, pp. 602–610.
[24] S. Nicholson, B. Milner, and S. Cox, "Evaluating feature set performance using the f-ratio and j-measures," in *Proc. EuroSpeech*, vol. 1, Rhodes, Greece, Sep. 1997, pp. 413–416.
[25] The National Gallery of the Spoken Word (NGSW). [Online] Available: http://www.ngsw.org/
[26] L. R. Rabiner and R. W. Schafer, *Digital signal processing of speech signals*. Englewood Cliffs, NJ: Prentice-Hall, 1978.
[27] J. Saunders, "Real time discrimination of broadcast speech/music," in *Proc. ICASSP*, vol. 2, Atlanta, GA, May 1996, pp. 993–996.
[28] E. Scheirer and M. Slaney, "Construction and evaluation of a robust multifeature speech/music discriminator," in *Proc. ICASSP*, vol. 2, Munich, Germany, 1997, pp. 1331–1334.
[29] M. Siegler, U. Jain, B. Raj, and R. M. Stern, "Automatic segmentation, classification and clustering of broadcast news audio," in *Proc. DARPA Speech Recognition Workshop*, Chantilly, VA, 1997, pp. 97–99.
[30] R. M. Stern, "Specification of the 1996 Hub4 broadcast news evaluation," in *Proc. DARPA Speech Recognition Workshop*, Chantilly, VA, 1997, pp. 3–6.
[31] SpeechFind, On-Line Search Engine for the National Gallery of the Spoken Word. [Online]. Available: http://SpeechFind.colorado.edu/
[32] TIMIT (recorded at Texas Instruments, transcribed at Mass. Inst. Technol.) Acoustic-Phonetic Continuous Speech Corpus. [Online]. Available: http://www.ldc.upenn.edu/
[33] G. Tzanetakis and P. Cook, "Multifeature audio segmentation for browsing and annotation," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, New Paltz, NY, Oct. 17–20, 1999, pp. w99-1–w99-4.
[34] S. Wegmann, P. Zhan, and L. Gillick, "Progress in broadcast news transcription at dragon systems," *Proc. ICASSP*, vol. 1, pp. 33–36, Mar. 1999.
[35] E. Wold, T. Blum, D. Keislar, and J. Wheaton, "Content-based classification, search and retrieval of audio," *IEEE Multimedia Mag.*, vol. 3, no. 3, pp. 27–36, 1996.
[36] U. Yapanel and S. Dharanipragada, "Perceptual MVDR-based cepstral coefficients (PMCCs) for noise-robust speech recognition," in *Proc. ICASSP*, vol. 1, Hong Kong, China, Apr. 2003, pp. 6–10.
[37] U. Yapanel and J. H. L. Hansen, "A new perspective on feature extraction for robust in-vehicle speech recognition," in *Proc. EuroSpeech*, Geneva, Switzerland, Sep. 2003, pp. 1281–1284.
[38] T. Zhang and C.-C. J. Kuo, "Audio content analysis for online audiovisual data segmentation and classification," *IEEE Trans. Speech Audio Process.*, vol. 9, no. 4, pp. 441–457, Jul. 2001.
[39] B. Zhou and J. H. L. Hansen, "Unsupervised audio stream segmentation and clustering via the Baysian information criterion," in *Proc. ICSLP 2000*, vol. 1, Beijing, China, Oct. 2000, pp. 714–717.
[40] B. Zhou and J. H. L. Hansen, "Speechfind: an experimental on-line spoken document retrieval system for historical audio archives," in *Proc. ICSLP*, vol. 3, Denver, CO, Sep. 2002, pp. 1969–1972.
[41] G. Zhou, J. H. L. Hansen, and J. F. Kaiser, "Nonlinear feature based classification of speech under stress," *IEEE Trans. Speech Audio Process.*, vol. 9, no. 3, pp. 201–216, May 2001.

Rongqing Huang (S'01) was born in China on November 12, 1979. He received the B.S. degree in electrical engineering from the University of Science and Technology of China (USTC) , Hefei, China, in 2002, and the M.S. degree in electrical engineering from the University of Colorado, Boulder, in 2004. He is currently pursuing the Ph.D. degree in electrical engineering at the University of Colorado.

From 2000 to 2002, he was a Research Assistant in the iFlyTek Speech Lab, USTC. From 2002 to 2005, he worked in the Robust Speech Processing Group, the Center for Spoken Language Research, University of Colorado. He was a Ph.D. Research Assistant in the Department of Electrical and Computer Engineering. In summer 2005, he was a Research Assistant with Motorola Research Laboratories, Schaumburg, IL. He is currently a Research Staff member in the Center for Robust Speech Systems, Human Language Technology Research Institute, University of Texas at Dallas, Richardson. His research interests include speech recognition and synthesis, machine learning and data mining, and digital signal processing and communication.



John H. L. Hansen (S'81–M'82–SM'93) received the Ph.D. and M.S. degrees in electrical engineering from the Georgia Institute of Technology, Atlanta, in 1988 and 1983, and the B.S.E.E. degree from Rutgers University, New Brunswick, NJ, in 1982.

In the fall of 2005, he joined University of Texas at Dallas (UTD), Richardson, where he is Professor and Department Chairman in Electrical Engineering, and holds an Endowed Chair in Telecommunications, and a joint appointment at the School of Behavioral and Brain Sciences (Speech and Hearing: Callier Center). At UTD, he established the Center for Robust Speech Systems (CRSS) which is part of the Human Language Technology Research Institute. Previously, he served as Department Chairman and Professor in the Department of Speech, Language, and Hearing Sciences (SLHS), and Professor in the Department of Electrical and Computer Engineering, at the University of Colorado, Boulder (1998–2005), where he cofounded the Center for Spoken Language Research. In 1988, he established the Robust Speech Processing Laboratory (RSPL) and continues to direct research activities in CRSS at UTD. His research interests span the areas of digital speech processing, analysis and modeling of speech and speaker traits, speech enhancement, feature estimation in noise, robust speech recognition with emphasis on spoken document retrieval, and in-vehicle interactive systems for hands-free human-computer interaction. He has supervised 33 Ph.D./M.S. thesis candidates.

Dr. Hansen, is serving as IEEE Signal Processing Society Distinguished Lecturer for 2005/2006, member of the IEEE Speech Technical Committee, and has served as Technical Advisor to U.S. Delegate for NATO (IST/TG-01), Associate Editor for the IEEE TRANSACTIONS on SPEECH AND AUDIO PROCESSING (1992–99), and Associate Editor for the IEEE SIGNAL PROCESSING LETTERS (1998–2000). He has also served as Guest Editor of the October 1994 Special Issue on Robust Speech Recognition for the IEEE TRANSACTIONS ON SPEECH AND AUDIO PROCESSING. He was recipient of the 2005 University of Colorado Teacher Recognition Award as voted by the student body, and author/coauthor of 205 journal and conference papers in the field of speech processing and communications, coauthor of the textbook *Discrete-Time Processing of Speech Signals*, (IEEE Press, 2000), and lead author of the report "The Impact of Speech Under 'Stress' on Military Speech Technology," (NATO RTO-TR-10, 2000). He also organized and served as General Chair for ICSLP-2002: International Conference on Spoken Language Processing, September 16–20, 2002.