ABSTRACT
This paper presents and compares methods for discrimination between speech from a broadcast audio device — like a television, radio, or GPS receiver — and live speech in the same acoustic environment. A solution to this discrimination problem has direct application wherever the audio from such a device interferes with voice recognition, verification, or transcription tasks. The methods and theory applied also have potential applications in multimedia and speaker segmentation, as well as in speaker verification. This paper presents a new use of the cepstral mean as an estimator of the linear time-invariant response of a “speaker” — either broadcast or live — over a relatively long time window. The problem is framed in terms of traditional speaker verification, but with two classes of speakers. This method is tested on five different data sets and the results compared for different feature sets, training methods, and window lengths.

Index Terms— Speech detection, cepstral mean, rich transcription, segmentation

1. INTRODUCTION
Considerable effort has gone into automated methods of modeling speech for voice activation and speaker verification applications. While this work builds upon the framework built up by speaker verification, the problem investigated here is fundamentally different. The goal of this work is to classify speech as belonging to one of two different sources in the same acoustic environment: speech from a television, radio, or other broadcast audio device and speech from a live person (or people). Access to a prior sample of speech (training data) for both classes and a single fixed-position microphone receiver in an relatively unchanged acoustic environment are assumed. A solution to this problem could be useful in a wide variety of applications including voice control of television, video games, or GPS-based navigation systems (where interfering broadcast speech is often or always present). Another potential application of this research is in preventing playback attacks [1] on speaker verification systems.

Section 2 presents a new use of the often intentionally discarded cepstral mean. The central hypothesis is that given a long enough sample of speech, different classes of speakers — referred to as broadcast and live — will be detectably different using a cepstral mean estimator of their linear time-invariant (LTI) response over a relatively long time window. Both mel-frequency cepstral coefficients (MFCCs) and real cepstra are tested and compared. Section 3 notes that broadcast and live speakers will each have time sections of multiple LTI responses and attempts to address this problem by modeling the broadcast speech class as \( N \) distinct speakers and the live speech class as \( M \) speakers. Section 4 formulates the decision test used for each long time window. The experimental design and five varied data sets are discussed in more detail in Section 5. Section 6 examines the results and compares the discriminative power of each feature set and training model. The paper concludes with a discussion of future work.

2. CEPSTRAL MEAN BASED MODELS
Subtracting the cepstral mean is commonly done to remove the LTI portion of speech in automatic recognition systems and/or to remove sensitivity to different microphones [2, 3]. But in this application the microphone and acoustic environment do not change between speakers. So instead of subtracting the information contained in the cepstral mean as in traditional speaker verification, this work tests its use as a feature. This use is justified with three assumptions. The first is that given a long enough speech sample, there will be sufficient variability to excite enough frequencies to characterize the LTI response of the speaker (either a live person or a broadcast audio device), their enclosure, and the passage to the receiving microphone. Also assumed is that these responses will dominate over the unreliable and changing room resonances. The third assumption is that the LTI response does not change significantly — either by a live talker moving around the room or by changing broadcast content, as discussed in Section 3 — over a testing decision window. Thus we hypothesize that the LTI components of live speech and broadcast speech in the same acoustic environment detectably differ and propose that they can be estimated using a cepstral mean over a relatively long time window.
This paper tests two different feature sets, mel-frequency cepstral coefficients (MFCCs) and real cepstra. While MFCCs are the standard feature set used in speaker verification, they are based on the human perception inspired “mel” scale [4, 5]. As this may be inappropriate for the detection of loudspeaker resonances (which are by design inaudible to humans), real cepstra are also tested. Real cepstra are defined as:

\[
\hat{x}_t = F^{-1}\{\log_{e} | F(x_t) |\},
\]

where \( F \) and \( F^{-1} \) are the Fourier transform and its inverse, respectively. For real cepstra, low-time liftering (low-pass windowing in the cepstral time domain) is used to remove the high time (cepstral quefrency) information, where unreliable and changing room resonances are dominant [6]. For both MFCCs and real cepstra the zeroth coefficient, a gain dependent term, is discarded.

### 3. MODEL TRAINING

This work also tests several strategies for modeling the cepstral representation of a speaker’s LTI room response. As a baseline, Gaussian mixture models (GMMs) are used to model the training data (transformed into a sequence of speech-like liftered cepstral frames) of a speaker class (either broadcast or live). However, a notable challenge in modeling the LTI response of a broadcast speech source is that frequently changing content often induces changes in the LTI response of that source. While these changes may be small in comparison to the contribution of the loudspeaker enclosure, path to the microphone, etc., changing the TV channel or radio station (or even an audibly subtle switch of studio acoustic environments) can significantly affect the overall LTI response. Similarly, for the live speech class there is the case of multiple talkers. This problem motivated the modeling of a single class’s varied training set as \( K \) distinct “speakers”, as labeled by a segmentation procedure. In order to segment a class’s training set the change decision criterion of [7] was used to give a measure of dissimilarity between two adjacent segments of speech (\( A \) and \( B \)):

\[
D_{A,B} = \frac{P(A \neq B)}{P(A \approx B)} = \frac{\sum_i P(a_i | \theta_A) + \sum_i P(b_i | \theta_B)}{\sum_i P(a_i | \theta_C) + \sum_i P(b_i | \theta_C)},
\]

where \( C = A \cup B \) and each element \( a_i \) or \( b_i \) is a single cepstral frame belonging to \( A \) or \( B \), respectively. Similarly \( \theta_A \) is the Gaussian model of the cepstral frames in set \( A \).

As preliminary work, this measure is calculated based on a sliding window of 200 frames of speech, which is moved over the whole training set. The \( K - 1 \) highest valued locations are chosen to denote changes in “speaker". Then each of the \( K \) contiguous time segments, each representing a single distinct LTI response, is modeled by a mixture of Gaussians.

### 4. DECISION THEORY

A decision between the live and broadcast speaker classes is made based on a time window of data, \( T \) cepstral frames in length. This window (indexed by \( i \)) is modeled by its cepstral mean \( \bar{c}_i \), which is the time average of its \( T \) individual cepstral frames. As there are \( N \) models for broadcast speech and \( M \) models for live speech (and no knowledge of the priors is assumed) the decision is formulated as a likelihood ratio test over the most likely models of each class. For this application this decision is formulated as:

\[
\lambda(\bar{c}_i) = \log \frac{\max_{1 \leq n \leq N} \left\{ P \left( \bar{c}_i \mid \theta_{\text{broadcast}}(n) \right) \right\}}{\max_{1 \leq m \leq M} \left\{ P \left( \bar{c}_i \mid \theta_{\text{live}}(m) \right) \right\}},
\]

where \( \theta_{\text{broadcast}}(n) \) is the \( n^{th} \) broadcast speech model, \( \theta_{\text{live}}(m) \) is the \( m^{th} \) live speech model, and \( P(\bar{c}_i | \theta) \) is the standard data likelihood for GMMs. Note that the numerator is identical in form to the numerator in a speaker verification decision, but that the denominator is also a maximum data likelihood over the models of a second class instead of the data likelihood of a universal background model. Also distinct in this decision is the use of a cepstral mean instead of a single (mean normalized) cepstral frame.

### 5. EXPERIMENTAL DESIGN

![Fig. 1. Block Diagram for Discrimination System.](image)
As shown in Figure 1, cepstral mean based discrimination between live speech and broadcast speech requires several steps, the details of which are discussed in this section. Single channel audio input is blocked into frames of 16 milliseconds. The voice operated switch (VOX) of [8] is used to eliminate frames which do not contain speech-like sound. Each frame is then pre-emphasized with a standard high-pass pre-emphasis filter with impulse response \( h[n] = \delta[n] - 0.98\delta[n] \) (where \( \delta[n] \) is a unit impulse). This is followed by a subtraction of the audio frame’s mean and the application of a von Hann window. The audio frame is then transformed into the cepstral domain, either using real cepstra (1) or mel-frequency cepstral coefficients as calculated by [9]. Delta cepstra are not used in this application. A lowpass lifter with a corner equal to one eighth the length of the cepstral frame is applied to real cepstra along with the elimination of the zeroth coefficient. The cepstral mean is then taken over a window of several frames. The segmented models are trained on speech-like audio frames which are pre-processed, transformed into cepstral frames, and liftered in the same manner as the testing audio but without a mean in time. The training procedure is conducted as discussed in Section 3. Gaussian mixture models are computed via expectation-maximization in the standard way, with diagonal covariance assumed. The likelihood ratio test is implemented as in (3), for each decision window. We divide both the broadcast and live speech-like frames into two parts for cross-validation.

A threshold computed from held-out data could be applied to the likelihood ratio of each decision window to make a hard decision between classes for each window. However, as different users of this system may have different goals when setting thresholds, we examine the receiver operating characteristic (ROC) curves with the goal of determining if the algorithm can produce enough discriminability for a given feature set, model type, and window length. ROC curves plot the rate of false positives (where live speech is mistakenly classified as broadcast) against the rate of true positives (where broadcast speech is correctly classified) for a given threshold choice, over the entire range of possible threshold choices [10]. The area under the ROC curve (AUC) was the figure of merit chosen for this experiment, as it is a single numeric measure often used by the machine learning community to compare many ROC curves [11]. For this application the AUC is also equivalent to the probability that a randomly chosen broadcast speech sample will have a smaller probability of being classified as live speech than a randomly chosen live speech sample [11]. An AUC of 1 represents a perfectly discriminating result, while an area of .5 would represent the results of random guessing. For this experimental set up the mean AUC is evaluated along with the 95% confidence interval. These are computed over the set of AUCs calculated on the four cross validations.

### Table 1. Description of Data Sets

<table>
<thead>
<tr>
<th>Set No.</th>
<th>Location</th>
<th>Broadcast Type</th>
<th>Length (sec)</th>
<th>Est. SNR</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Office</td>
<td>TV News</td>
<td>341</td>
<td>114</td>
</tr>
<tr>
<td>2</td>
<td>Office</td>
<td>TV News</td>
<td>352</td>
<td>227</td>
</tr>
<tr>
<td>3</td>
<td>Home</td>
<td>Movie</td>
<td>329</td>
<td>412</td>
</tr>
<tr>
<td>4</td>
<td>Lab</td>
<td>Podcast News</td>
<td>445</td>
<td>631</td>
</tr>
<tr>
<td>5</td>
<td>Car</td>
<td>Radio News</td>
<td>321</td>
<td>40</td>
</tr>
</tbody>
</table>

### Table 2. Preliminary Results: 1/4 Second Testing Window, \( K = G = 1 \). (An area under the ROC curve of 1.00 indicates perfect experimental performance.)

<table>
<thead>
<tr>
<th>Set No.</th>
<th>Area Under the ROC Curve Real Cepstra</th>
<th>Area Under the ROC Curve MFCCs</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.00±0.00</td>
<td>1.00±0.00</td>
</tr>
<tr>
<td>2</td>
<td>1.00±0.00</td>
<td>0.97±0.01</td>
</tr>
<tr>
<td>3</td>
<td>0.94±0.03</td>
<td>0.87±0.01</td>
</tr>
<tr>
<td>4</td>
<td>1.00±0.00</td>
<td>0.98±0.03</td>
</tr>
<tr>
<td>5</td>
<td>0.98±0.01</td>
<td>0.93±0.03</td>
</tr>
</tbody>
</table>

To the best of our knowledge there is no published database available for testing a system to distinguish broadcast speech from live speech in the same acoustic environment. Therefore we recorded several minutes of broadcast and live speech data in five different acoustic environments, as briefly described in Table 1. The third data set contained multiple live speakers (including a man, a woman, and several children), while the other data sets each contain a single male speaker. All data sets contain relatively varied broadcast content. The estimate of signal to noise ratio (SNR) was calculated by a ratio of the average energy in the speech-like frames over the average energy in the non-speech frames. As non-speech-like sound will count as noise, this method is likely to overestimate the amount of noise. However it provides a consistent basis for quantitative comparison between the different data sets.

### 6. Preliminary Results

The experiments were designed to test and compare the effectiveness of the two feature sets discussed in Section 2 and two methods of modeling discussed in Section 3. Gaussian mixtures with \( G = \{1, 2, 4, 8\} \) Gaussians in each mixture were tested and compared to segmented speaker single Gaussian models with the same number of parameters \( K = \{1, 2, 4, 8\} \) speakers to represent each class, with each speaker modeled by a single Gaussian). Window lengths from 160ms to 10 seconds were tested. The results of this experiment, with \( G = K = 1 \) and a testing window length of 1/4 second, are detailed in Table 2. This window length was chosen for
the purpose of contrasting the two feature sets’ performance without saturation. Both feature sets attain perfect classification on all data sets given window lengths greater than 3 seconds. The simplest single Gaussian and single speaker models are displayed here because they outperform more complex higher order models, for both feature sets and all window sizes.

Given these results we conclude that the LTI components of live speech and broadcast speech in the same acoustic environment detectably differ and that they can be estimated using cepstral means over relatively long time windows. Furthermore these results indicate that liftered real cepstral means may be a better estimator than liftered MFCC means for this application. We note that decreased SNR may have a negative effect on the performance of real cepstral means as an estimator. While models with many Gaussians per mixture are normally used to describe speech data, in this application single Gaussian models outperform mixtures of Gaussians and attempts to build multiple speaker models via segmentation. This is most likely due to having insufficient data in a few minutes of speech data to accurately estimate such large numbers of model parameters. Also notable is that the cepstral coefficients of a speech sample in this application generally have a unimodal but skewed distribution. Such data may be better modeled by a distribution which could take into account this skew, but with fewer parameters.

7. CONCLUSION AND FUTURE WORK

In this paper we presented a new application — discrimination between broadcast and live speech in the same acoustic environment — and modified traditional speaker verification techniques to take advantage of the differences between the applications. We showed that real cepstral means over relatively long time periods can be used as an estimator of the LTI responses of live and broadcast speech, that modeling these responses over training data sets can be done, and that these two classes are perfectly discriminable for all data sets tested, given three second or longer decision windows.

While this initial study shows promising results, there are many potential directions for future work. Longer and more varied data sets are needed to fully test our hypotheses. Further analysis of the effectiveness of cepstral mean modeling of LTI responses in low SNR environments and over shorter time windows is needed. The training of live and broadcast speech models is another direction with great potential. This could include further study of the best distribution and modeling technique for this type of data. Additionally, much careful work has been done in the area of audio segmentation which if applied carefully to this application may improve performance on data where the source’s LTI response varies.

8. ACKNOWLEDGEMENTS

The authors would like to acknowledge helpful discussions with Prof. Bishnu Atal, Pascal Clark, Brian King, and other colleagues at the Department of Electrical Engineering at the University of Washington. We also thankfully acknowledge Nima Mesgarini and Prof. Shihab Shamma of the University of Maryland for their use of their VOX to pre-process our data.

9. REFERENCES


