AUDIO-BASED CLASSIFICATION OF SPEAKER CHARACTERISTICS

Promiti Dutta and Alexander Haubold

Columbia University, New York, NY
{pd2049,ah297}@columbia.edu

ABSTRACT

The human voice is primarily a carrier of speech, but it also contains non-linguistic features unique to a speaker and indicative of various speaker demographics, e.g. gender, nativity, ethnicity. Such characteristics are helpful cues for audio/video search and retrieval. In this paper, we evaluate the effects of various low-, mid-, and high-level features for effective classification of speaker characteristics. Low-level signal-based features include MFCCs, LPCs, and six spectral features; mid-level statistical features model low-level features; and high-level semantic features are based on selected phonemes in addition to mid-level features. Our data set consists of approximately 76.4 hours of annotated audio with 2786 unique speaker segments used for classification. Quantitative evaluation of our method results in accuracy rates up to 98.6% on our test data for male/female classification using mid-level features and a linear kernel support vector machine. We determine that mid- and high-level features are optimal for identification of speaker characteristics.

Index Terms— audio signal processing, feature extraction, MFCC, LPC, classification, gender, ethnicity

1. INTRODUCTION

Searching through vast amounts of spoken audio collections is an arduous task without the availability of search cues. Audio transcripts generated by Automatic Speech Recognition (ASR) systems provide good content search cues, albeit imperfect coverage and varying accuracy, especially for salient key terms [1,2]. Search for content can be improved significantly through re-ranking or filtering speech segments by known speaker characteristics.

In this paper we identify and evaluate classifiers for three characteristics: gender, nativity (native vs. non-native English), and ethnicity (African-American, Asian, Caucasian, Hispanic, South-east Asian). Using a large dataset of 2786 manually annotated speech segments from student presentation videos, we evaluate and train on various low-, mid-, and high-level feature classifiers on the detection of voice characteristics (Figure 1). Through experimentation, we observe that low-level features are significantly less effective in determining characteristics than mid-level features. For gender classification, we achieve an accuracy of 67.3% using low-level signal-based features. Tzanetakis et. al. report similar results (76%) on TRECVid 2003 data; however their low-level features are computed on 20 ms windows, while we use 10 ms windows [3]. Our results for mid-level statistical features show significant improvement, leading to an overall accuracy of 90.1%-98.6% over varying speech window sizes.

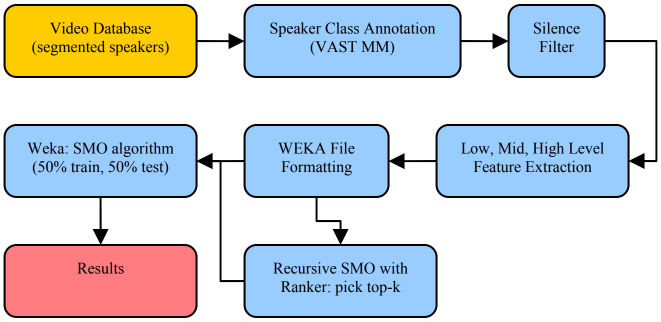

Video Database (segmented speakers)
Speaker Class Annotation (VAST MM)
Silence Filter
Low, Mid, High Level Feature Extraction
WEKA File Formatting
Weka: SMO algorithm (50% train, 50% test)
Results
Recursive SMO with Ranker: pick top-k

Figure 1: Overview diagram of processing steps in our approach.

2. DATA SET

Our dataset includes student final project presentation videos from a large university-level engineering design course with more than 150 students per semester. Each presentation team is comprised of 5 – 6 students who take turns presenting their team's project during a midterm and a final period in the semester. Our video data spans 5 years.

We perform data annotation to establish ground truth using the VAST MM *Video Audio Structure Text Multimedia* system (Figure 2) [4]. The VAST MM browser displays audio and visual cues, which are useful for distinguishing speaker segments. In an indexing step, the VAST MM indexing tool performs several content analysis processes, including automatic speaker segmentation based on Mel Frequency Cepstral Coefficient (MFCC) features and the Bayesian Information Criterion (BIC) [5]. Using the tool, we listen to and view short video clips from each speaker segment to correctly annotate each with appropriate classifications. Each speaker segment is classified according to gender, ethnicity, and familiarity of spoken English.

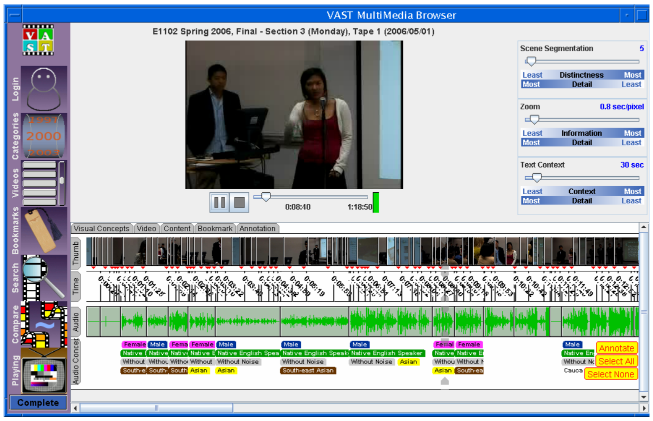


Figure 2: VAST MM browser used to annotate speaker segments. Visual cues (key frames and streaming video) and audio signal are displayed in the user interface for ease of annotation.

Table 1 summarizes the sample sizes of the annotated data set: we have annotated over 76.4 hours of audio with 2786 unique speaker segments. Each audio speaker segment is extracted from the original video for further analysis.

In a preprocessing step, audio speaker segments are filtered for silence. This step is crucial for removing a signal, which would otherwise act as a similarity between speaker segments from different classes. Because the original video recordings were made with wired and wireless analog microphones, silent pauses in the audio track are practically low-amplitude noise. Their numerical representation as MFCC features is substantially different from actual speech: the zeroth MFCC feature, commonly referred to as a representation of signal amplitude, deviates most; higher order MFCCs also reflect a significant difference due to the high frequency inherent to noise. We apply a simple heuristic, which computes the absolute maximum amplitude A for a given speaker segment, and filters out any short fixed audio sample window (256 samples) which does not pass a threshold measured as an empirically determined fraction of A.

Key characteristics of the audio data include varying audio quality between student presentations. This is largely due to different microphones that were used over the five-year course recordings. Also affecting audio quality is an individual speaker's use of the microphone, such as placement with respect to speaker (hand-held vs. on-stand) and presenter's activity (rigid pose vs. constant shifting).

We notice a skew in the distribution between certain annotation classes. Specifically, in the engineering school we observe a 3:1 ratio of male to female students. Similarly, we find fewer speakers in some ethnic classes (African Americans and Hispanics) than others (Asians, Caucasians, and Southeast Asians). To avoid a bias due to unequal sample sizes, we down-sample the data set to comparable class sizes for classification.

Table 1: Summary of classification for data set.

	Class	# of Segments	Time (hr)
Ethnicity	African-American	101	2.36
	Asian	776	20.15
	Caucasian	1233	33.34
	Hispanic	80	2.00
	South-east Asian	295	7.74
Gender	Male	1865	51.86
	Female	692	16.23
Spoken English	Native	2197	58.81
	Non-native	327	8.53

3. FEATURE EXTRACTION

We extract low-, mid-, and high-level features from each audio speaker segments for varying time-intervals. Low-level features are signal-based; mid-level features are statistical aggregates of low-level features; and high-level features include phonemes in addition to mid-level features.

3.1. Low-level: Signal-level

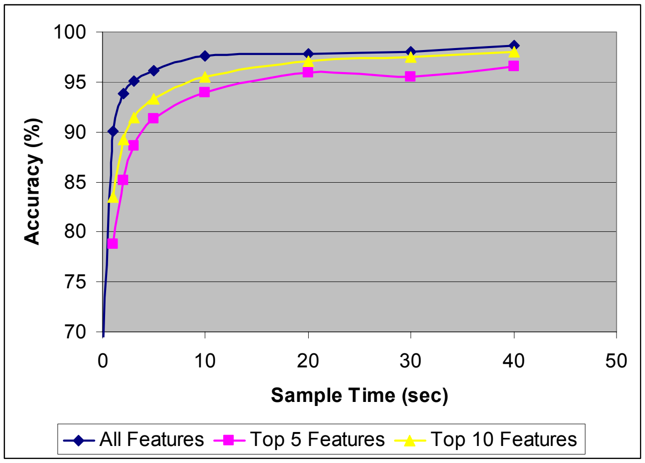
Low-level features include 13 MFCCs, 13 Linear Predictive Coefficients (LPCs), and 6 distinct spectral features for a total of 32 distinct features from each 256-sample window (~0.01 sec in a 22 kHz sampled signal). The 13 MFCCs are a representation of the short-term power spectrum of a sound. LPCs analyze the speech signal by estimating formants, removing their effects from the speech signal, and estimating the intensity and frequency of the remaining buzz. The six spectral features include energy entropy block, short time energy, zero-crossing rate, spectral roll off, spectral centroid, and spectral flux [6].

3.2. Mid-level: Statistical Aggregates from Signal Level

Mid-level features are statistical aggregates of the aforementioned 32 low-level features on longer samples. The low-level features underlie a Gaussian distribution with mean, μ, and variance, σ. We model the aggregate of low-level MFCC and LPC features by their mean and covariance. The covariance matrices for MFCC and LPC are symmetrical; we only use the covariance values from the upper triangular matrix and the diagonal for a total of 91 values for MFCCs and LPCs, respectively. We include 13 MFCC means, 13 LPC means, and respective statistical measures for the 6 spectral features [6]. The complete feature vector for mid-level features contains 214 features.

3.3. High-level: Semantic Level

High-level feature vectors are derived from mid-level features. We include 12 additional features derived from phonemes for a total of 226 features (91 MFCC cov, 13 MFCC mean, 91 LPC cov, 13 LPC mean, 6 spectral



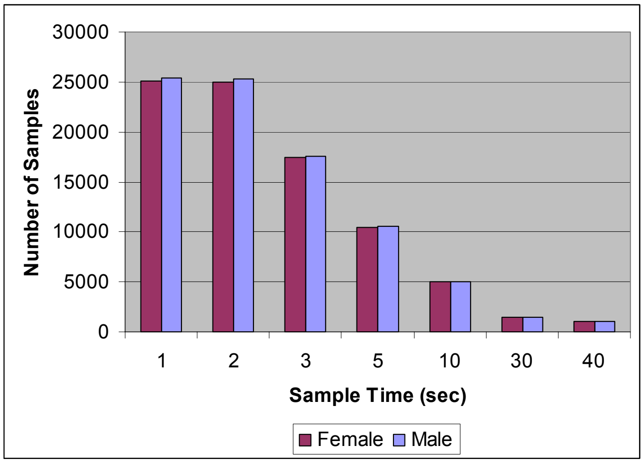


Figure 3 (left): Male/female classification accuracy for varying sampling time lengths. The 256-sample window (~0.01 sec) using low-level samples is our baseline accuracy (67.3%). Mid-level features (1 sec – 40 sec) exhibit significant increasing classification accuracy. The same trend is apparent after applying classification with the top 5 and 10 features selected by recursive feature selection. **Figure 4 (right):** Distribution of non-overlapping sample sizes for male/female speaker segments at different sampling time durations.

features, and 12 phonemes). We apply phoneme extraction to generate a frequency list of occurring phonemes in the audio signal. We apply our approach [7] to identify a selection of monophthongs, diphthongs, and fricatives (/AA/, /AE/, /AH/, /AO/, /EH/, /ER/, /IH/, /IY/, /S/, /SH/, /UH/, /UW/). This heuristic method models the vocal tract using an autoregressive model of the speech signal in which the peaks of the frequency response correspond to resonant frequencies of the vocal tract (formants). The closest matching phoneme is determined by the Euclidean distance of a weighted difference between model and computed values by using a table of expected frequency values for formants F1, F2, and F3.

4. CLASSIFICATION AND FEATURE SELECTION

Classification is performed using the Sequential Minimal Optimization (SMO) algorithm in Weka [8]. SMO is a computationally simpler method to compute the support vector machine (SVM) quadratic programming (QP) optimization problem without extra matrix storage and without using numerical QP optimization steps. We use a linear kernel unless otherwise noted for the SMO. The output equation for a linear SVM (Equation 1) defines w as the normal vector to the hyperplane, x as the input vector, and u as the separating hyperplane. The linear kernel identifies the optimal separating hyper-plane between the distributions by maximizing the margin m (Equation 2) using training examples. Prediction is performed on the test set. To avoid classification biases, cross-validation is obtained for the experiments using a 50% split for training and test sets.

$$u = \vec{w} \cdot \vec{x} - b \qquad \text{(Equation 1)}$$

$$m = \frac{b}{\|w\|_2} \qquad \text{(Equation 2)}$$

Additionally, we perform feature selection using the "SVMAttributeEval" method in Weka. "SVMAttributeEval" evaluates the weight of an attribute by using a linear SVM. The "Ranker" search method ranks each feature by the square of the weight assigned by the SVM from the "SVMAttributeEval" method. The selected features are used for classification to test the overall prediction of the dataset.

5. EXPERIMENTS AND RESULTS

5.1. Low-level: Signal-level

We apply low-level features to non-overlapping 256-sample windows (~0.01 sec) for gender classification (male/female). Male samples sizes are down-sampled to adjust for any classification bias due to mismatching sample sizes between the two classes. In total, we have 105,106 male speaking auditory samples compared to the 94,555 female speaking samples. The linear kernel SMO achieves 67.3% classification accuracy (Figure 3), consistent with an accuracy of 76% on 0.02 sec sample windows in [4].

5.2. Mid-level: Statistical aggregates from signal level

5.2.1. Varying Sampling Times

We extract mid-level features for several non-overlapping sampling intervals: 1, 2, 3, 5, 10, 20, 30, and 40 seconds (Figure 4). Monologues longer than 40 seconds by a given speaker are rare in our dataset. We down-sample the male samples to create equal distribution between male and female samples for classification.

We perform classification using all 214 features to determine the efficacy of mid-level features at varying sampling intervals. Classification accuracies range between 90.1 – 98.6% (Figure 3) where accuracy is logarithmically related to sample time. A 10-second time interval provides a reasonable baseline for analysis on high-level features.

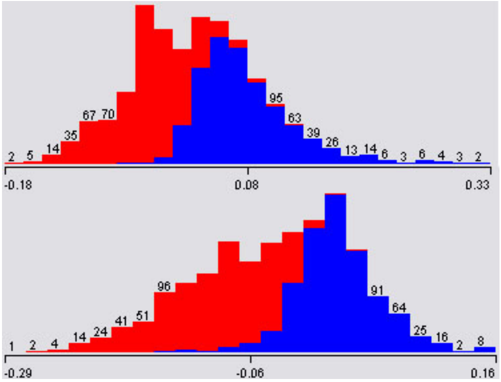


Figure 5: Additive histogram of male samples shown in blue; female in red. (top) MFCC covariance 2_7. (bottom) MFCC covariance 3_8.

5.2.2. Feature Selection

Each feature vector contains 214 features for mid-level analysis. The use of excessive features can result in over-fitting. To determine whether the data was over-fitted, we perform feature selection to identify the 5 and 10 most significant features for our classification for each sampling interval (1, 2, 3, 5, 10, 20, 30, and 40 seconds). We determine that using fewer features provides comparable classification accuracy and is less computationally expensive (Figure 3).

For each of the sampling intervals, we obtain a distinct group of top 5 and 10 significant. We note that certain features are common to each sampling interval. Specifically, there are two MFCC co-variances that rank as the top two features for all classification performed with mid-level features. The additive histogram contains two very distinct Gaussians for males and females with very different means μ for these two MFCC co-variances (Figure 5).

5.3. High-level: semantic level

5.3.1. Spoken English Experiment

In the spoken English experiment, we classify native English speakers versus non-native English accented speakers. Sample size is 2700 male and 2700 female feature segments. We obtain a 73.5% classification accuracy.

It is possible that the classifier is confounded by demographical data. We perform additional experiments in which we create sub-groups for classification, i.e. African American native English speaker versus African American non-native English speaker. Classification accuracy for these smaller groups rises to 80% and greater for each of the 5 smaller subgroups created.

5.3.2. Demographics Experiment

The demographics experiment is a multi-class classification task amongst five groups: African Americans, Asians, Caucasians, Hispanics, and South-east Asians. We sample each group to contain 600 samples of non-overlapping 10-second sampling windows. We obtain 48.5% classification accuracy using an empirically determined 5th degree polynomial kernel. A linear kernel did not provide effective classification accuracies.

The demographic classification may be confounded by inclusion of both native and non-native English speakers in the respective groups. We remove this bias by creating groups based on native and non-native speakers and their respective demographics class. This increases classification accuracy to approximately 64.5% accuracy for each class. The classification confusion matrix indicates the similarity between the Asian and South-east Asian groups, suggesting that better accuracy may be obtained by combining these two groups. A similar association is observed with the African American and Hispanic groups as well. We note that these results are significant compared to the probabilistic 20% accuracy achieved by random guessing.

6. CONCLUSIONS

This paper presents a survey of the different levels of features that can be applied to classification of speech. We demonstrate that low-level features perform poorly for classification since audio sampling is too short and therefore not representative of characteristic traits for the classification classes. We show that mid- and high-level features perform significantly better, because higher order features more closely correspond to human perception of auditory characteristics. A human can identify characteristics best with ample amount of information (longer speech segments) rather than short samples of speech. The main disadvantage with the use of these types of features is the requirement of longer audio segments. However given the domain of presentation and lecture videos, these audio segments are aptly available and thus are applicable for effective audio search methods for large multimedia collections. We propose further investigation into the high-level feature domain by through the exploration of additional phonemes as well as semantic (vocabulary) usage.

7. REFERENCES

[1] B. Matthews, U. Chaudhari, and B. Ramabhadran, "Fast Audio Search Using Space Modeling," ASRU '07, Kyoto, Japan.

[2] C. González-Ferreras and V. Cardeñoso-Payo, "A System for Speech Driven Information Retrieval," ASRU '07, Kyoto, Japan.

[3] G. Tzanetakis, M.-Y. Chen, "Building Audio Classifiers for Broadcast News Retrieval," WIAMIS '04, Lisbon, Portugal, 2004.

[4] A. Haubold, J.R. Kender, "VAST MM: Multimedia Browser for Presentation Video", CIVR '07, Amsterdam, The Netherlands.

[5] A. Haubold, J.R. Kender, "Augmented segmentation and visualization for presentation videos," MM '05, Singapore.

[6] T. Giannakopoulos, "Some Basic Audio Features," Matlab File Exchange, March 16, 2008.

[7] A. Haubold, J.R. Kender, "Alignment of Speech to Highly Imperfect Text Transcriptions", ICME '07.

[8] I.H. Witten and E. Frank, "Data Mining: Practical machine learning tools and techniques," 2nd Edition, Morgan Kaufmann, San Francisco, 2005.