SPEAKER IDENTIFICATION USING A MICROPHONE ARRAY AND A JOINT HMM WITH SPEECH SPECTRUM AND ANGLE OF ARRIVAL

Jack W. Stokes, John C. Platt, and Sumit Basu

Microsoft Research, One Microsoft Way, Redmond, WA 98052, USA
{jstokes, jplatt, sumitb}@microsoft.com

ABSTRACT

In this paper, we present a speaker identification algorithm for a microphone array based on a first-order joint Hidden Markov Model (HMM) where the observations correspond to the angle of arrival of the speech and the speech spectrum. The goal of the research is to investigate whether including angle of arrival information improves the speaker identification error rates compared to an algorithm based on the speech spectrum only. The spectral model consists of a Gaussian Mixture Model (GMM) using Multiple Discriminant Analysis (MDA) coefficients, and the angle model includes a separate histogram for each participant. The convergence time of the joint HMM is improved by estimating the GMM for each of the meeting participants prior to the start of the meeting and initializing each participant's spectral GMM in the joint HMM to the pretrained parameter values. The performance of the algorithm is analyzed from data collected during live meetings recorded using an eight element, circular microphone array. For meetings where the participants are stationary, the results show significant improvement over a single channel speaker ID algorithms based on spectrum only.

1. INTRODUCTION

Meeting transcription is a desirable, but extremely challenging, feature of automated meeting systems [1, 2, 3]. A preliminary step to enabling meeting transcription is identifying who is speaking for each frame of audio data. This problem is commonly referred to as speaker diarization or speaker identification. By first preprocessing the audio signal with a speaker identification algorithm, individual speech recognition models can then be used on the segmented speech pertaining to each of the meeting participants.

Speaker identification based on a single channel of audio data has been an active area of research [4]. In meetings with several people, close-talk microphones are not practical: these meetings can be recorded with microphone arrays. In the past, several papers [5, 6, 7] have suggested using a microphone array to preprocess the audio signal in order to improve speaker identification as compared to processing a single channel speech signal. Other papers have proposed using the angle of arrival from a microphone array to segment (but not identify) the speaker [8].

This paper uses a hybrid between the two approaches: we create a speaker identification system based on a Hidden Markov Model (HMM) where the observations form a joint distribution over the angle of arrival of the speech and the speech spectrum produced by a beamformer. This work extends previous work [9] with joint angle/spectrum models, by operating on the beamformed output, rather than deriving a spectrum from close-talking microphones.

Similar to [8], the angle of arrival from the microphone array provides a very strong hint for segmenting the speech between speakers. For the case where the participants do not move during the meeting and where they are not located in close, angular proximity to one another, clustering based purely on the angle of arrival may yield sufficient segmentation results. However, clustering does not determine the identity of the person speaking and fails when two individuals are located next to each other or one directly behind the other. In addition, the angle estimate can also be distorted by reflections from objects within the meeting room (e.g. white-boards, laptop computers, etc.). Single channel speaker identification algorithms based on the speech spectrum can produce high error rates unless significantly smoothed during post processing. By combining both the angle information from the microphone array with the beamformed speech spectrum, the proposed speaker identification algorithm is able to significantly improve speaker identification error rates by making the decision based on the speech spectrum for data coming from only a single direction. This new speaker identification algorithm provides extremely promising results on audio data captured from an 8 element, circular microphone array for the case where the participants are stationary.

The paper is organized as follows. In the following section, we provide a high level description of the joint Hidden Markov Model employed in this paper. The speaker identification algorithm based on the joint HMM is presented in detail in section 3. Finally, numerical results from three experimental meetings, and conclusions are given at the end of the paper.

2. JOINT HIDDEN MARKOV MODEL

In this section and the following section, we describe the first-order joint HMM following the notation and derivation given in [10]. An HMM is a probabilistic model which is governed by a set of states and an associated state transition probability matrix A with elements a_{ij}. At any given time step t, the current state depends solely on the previous state at time $t-1$ and cannot be observed. The time index ranges from $t = 1, \cdots, T$ where T is the time index for the final observation. For the case of speaker identification, each hidden state i represents one unknown person speaking and ranges from $i = 1, \cdots, N$ given N participants in the meeting. In this work, we assume that the participants are known at the start of the meeting. In a standard HMM, the hidden state produces a single observation. In this paper, we employ a joint HMM, illustrated in figure 1, where the hidden state generates two observations: the angle of arrival of the speech from the current person speaking which is detected by the microphone array, and the spectrum of the speech output from the microphone array's beamformer. The angle of arrival is estimated using a time delay of arrival algorithm [11]. The probability of observing the *spectral* observation and the *angle* observation at time t for speaker i are given by $b_{s,i}(o_{s,t})$ and $b_{a,i}(o_{a,t})$, respectively.

1-4244-0367-7/06/$20.00 ©2006 IEEE

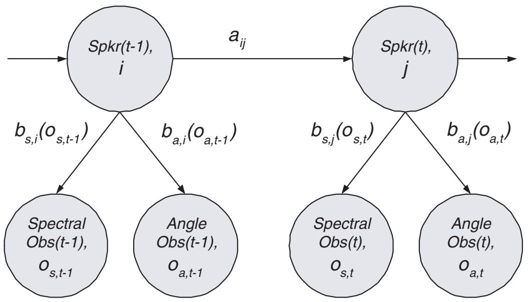

Fig. 1. Joint HMM for speaker identification with speech spectrum and angle observations.

3. SPEAKER IDENTIFICATION ALGORITHM

Next, we describe the audio system and derive the joint HMM algorithm. Multichannel audio data is captured from the eight, unidirectional microphones, and each of the time domain signals is converted to the frequency domain using the modulated complex lapped transform (MCLT) [12]. Alternatively, the time to frequency domain conversion can be computed using other transforms such as the fast Fourier transform (FFT). The microphone array processes the multichannel MCLT stream and produces a single channel, beamformed output at each frame. This beamformed output is used to compute the spectral observation for the HMM. Furthermore, for those frames containing speech, as determined by a voice activity detector, the angle of arrival is also estimated as part of the microphone array algorithm

3.1. SPECTRAL PROCESSING

In the microphone array hardware, the audio data is sampled at 16 kHz with 20 msec frames. Typical speaker identification algorithms process a single frame (e.g. 20 msec) at a time. In this algorithm, the magnitude of the MCLT spectrum from twenty-four, 20 msec frames of audio data are aggregated. Only the 220 bands above the lower 50 Hz for each spectral frame are processed and aggregated. The choice of using 24 aggregated frames is determined by listening tests of humans being able to identify a speaker based on 500 msec of speech. In order to remove the spectral tilt effects from the microphones, the MCLT is further preprocessed using a liftering technique [13]. If twelve or more of the twenty-four frames contain valid speech, the aggregated spectral vector is projected into a lower subspace of dimension D using multiple discriminant analysis (MDA). The decision to use MDA coefficients is based on the results from table 1 for a single channel speaker ID algorithm using only spectral acoustic data. In table 1, we investigate the performance of MDA coefficients versus Mel Frequency Cepstral Coefficients (MFCCs). In addition, we also compare K nearest neighbor (KNN) and Gaussian Mixture Models (GMMs) for the classification method. The results in table 1 do not include post processing to smooth, or filter, the outputs to remove spurious incorrect speaker decisions. Post processed smoothing significantly reduces error rates for single channel speaker identification algorithms. For example, smoothing drops the error rate from 33.5% to 15.3% for the case of MDA with 30 coefficients and $K = 9$ KNN classification. The MDA projection coefficients are found by maximizing the between class distances while simultaneously minimizing the within class distances and were trained on the TIMIT database. In this paper, we use a separate GMM to model the spectral observations from each speaker for the joint HMM. The D dimensional spectral feature vector at time t is denoted by $o_{s,t}$.

Feature Vector	Classification Method	Error Probability
30 MDA Coefs	KNN (K = 9)	33.5%
13 MFCCs	GMMs (N = 10)	39.5%
30 MDA Coefs	GMMs (N = 10)	30.1%
40 MFCCs	KNN (K = 9)	35.2%

Table 1. Speaker ID error rates for various algorithms.

A key step in the algorithm is pretraining the spectral GMMs for individual speakers prior to the start of the meeting. At the beginning of the meeting, the GMMs for the separate states representing each of the known participants in the meeting are initialized to the original, pre-trained GMMs corresponding to that person. Each speaker's GMM is then adapted over the coarse of the meeting. For meeting participants without pretrained GMMs, a short training session can be conducted at the start of the meeting or those participants can be identified as unknown participants.

3.2. ANGLE OF ARRIVAL PROCESSING

While the speech spectrum for each speaker is modelled with a GMM, the angle of arrival is modelled by a separate histogram for each speaker. The microphone array produces an angle of arrival estimate in the range of 0 to 360 degrees. The number of valid speech frames in the twelve most recent raw angle measurements are averaged to smooth the angle estimate. This filtered angle estimate is then quantized into a histogram with N_a bins. The angle observation $o_{a,t}$ represents the index of the histogram associated with the microphone array's angle of arrival estimate at time t.

3.3. JOINT E STEP

The joint HMM model parameters are trained using the EM (Expectation - Maximization) algorithm. In order to highlight the differences between the joint HMM and the standard HMM presented in [10], we next give the details for the joint EM algorithm. In the E Step, the model parameters for the joint HMM λ_h, the angle model λ_a, and the spectral model λ_s are fixed and the responsibility vectors are updated. The first step is to compute the forward variables

$$\alpha_i(t) = p(O_1 = o_1, \cdots, O_t = o_t, Q_t = i|\lambda_h, \lambda_a, \lambda_s) \quad (1)$$

for $t = 1, \cdots, T$ using the forward algorithm and the joint observation vector at time t composed of the angle and spectral observations $o_t = \{o_{a,t}, o_{s,t}\}$. Similarly, the backward algorithm computes the backward variables $\beta_i(t)$ for $t = 1, \cdots, T$ where

$$\beta_i(t) = p(O_{t+1} = o_t, \cdots, O_T = o_T|Q_t = i, \lambda_h, \lambda_a, \lambda_s). \quad (2)$$

After computing $\alpha_i(t)$ and $\beta_i(t)$, the joint responsibility vectors can now be computed as

$$\gamma_i(t) = p(Q_t = i|O, \lambda_h, \lambda_a, \lambda_s) = \frac{\alpha_i(t)\beta_i(t)}{\sum_{j=1}^{N} \alpha_j(t)\beta_j(t)} \quad (3)$$

and the probability of speaker i talking at time t and speaker j talking at time $t + 1$ as

$$\xi_{ij}(t) = \frac{\gamma_i(t)a_{ij}b_j(o_{t+1})\beta_j(t+1)}{\beta_i(t)} \quad (4)$$

for $i = 1, \cdots, N$, $j = 1, \cdots, N$ where $b_j(o_{t+1})$ is the joint output probability at time $t + 1$.

Finally for the spectral GMMs, we compute the responsibility vectors for each Gaussian component in the GMM as

$$\gamma_{il}(t) = \frac{\gamma_i(t) c_{il} b_{s,il}(o_{s,t})}{b_{s,i}(o_{s,t})} \tag{5}$$

where c_{il} is the GMM mixing coefficient for the l^{th} Gaussian of speaker i's mixture.

3.4. JOINT M STEP

In the M Step, the responsibility vectors are fixed and the parameters of λ_h, λ_s and λ_a are updated. To update the spectral model λ_s, the mixing coefficients c_{il}, mean μ_{il}, and the diagonal covariance matrix Σ_{il} for the l^{th} Gaussian of speaker i's mixture are updated as

$$c_{il} = \frac{\sum_{t=1}^{T} \gamma_{il}(t)}{\sum_{t=1}^{T} \gamma_i(t)} \tag{6}$$

$$\mu_{il} = \frac{\sum_{t=1}^{T} \gamma_{il}(t) o_{s,t}}{\sum_{t=1}^{T} \gamma_{il}(t)} \tag{7}$$

$$\Sigma_{il} = \frac{\sum_{t=1}^{T} \gamma_{il}(t)(o_{s,t} - \mu_{il})^2}{\sum_{t=1}^{T} \gamma_{il}(t)}. \tag{8}$$

The algorithm currently assumes a diagonal covariance for each spectral GMM although a full covariance could also be used given enough data. Furthermore, we require all variances on the diagonal to be greater than or equal to some minimum value, $\Sigma_{il}(m,m) >= \sigma^2_{min}$.

Next, the joint HMM model parameters are updated including the initial probabilities $\pi_i = \gamma_i(1)$ and the transition matrix coefficients

$$a_{ij} = \frac{\sum_{t=1}^{T-1} \xi_{ij}(t)}{\sum_{t=1}^{T-1} \gamma_i(t)} \tag{9}$$

for $i = 1, \cdots, N$, $j = 1, \cdots, N$. To compute the joint output probabilities, we must first compute the output probabilities of the angular histogram for each bin k corresponding to the quantized angle observations

$$b_{a,i}(k) = \frac{\sum_{t=1}^{T} \delta_{o_{a,t}, v_k} \gamma_i(t)}{\sum_{t=1}^{T} \gamma_i(t)} \tag{10}$$

and for the spectral data

$$b_{s,i}(o_{s,t}) = p(O_t = o_{s,t} | Q_t = i) = \sum_{l=1}^{M} c_{il} b_{s,il}(o_{s,t}) \tag{11}$$

where M is the number of mixtures in the GMM. The joint output probability can then be computed as

$$b_i(o_t) = (1/D) b_{s,i}(o_{s,t}) b_{a,i}(o_{a,t}). \tag{12}$$

3.5. INITIALIZATION

Prior to running the joint EM algorithm, the transition matrix A is initialized so that the current frame of speech most likely comes from the same person who spoke the previous frame. We can achieve this by initializing the diagonal elements of A with a probability close to one and the off-diagonal probabilities of equal value so that the sum of the transition probabilities out of a particular state equal one. For example, we set $a_{ii} = 0.95$ and $a_{ij} = 0.05/(N-1)$. Next, we initialize the probability of the observation sequences for the angle estimates to be uniform across the N_a bin histogram, $b_{a,i}(o_{a,0}) = 1/N_a$. The output probabilities for the spectral data based on the spectral GMMs are initialized as

$$b_{s,i}(o_{s,0}) = \sum_{l=1}^{M} c_{il} b_{s,il}(o_{s,0}). \tag{13}$$

With the spectral and angle output probabilities initialized, the joint output probability can also be initialized as in (12) with $t = 0$. Finally, the initial probabilities, $\pi(i)$, are initialized so that they are approximately uniformly distributed plus a small amount of random noise.

3.6. SPEAKER IDENTIFICATION

Once the joint HMM has been trained, the estimated state sequence identifying which person spoke for each frame can now be found from the spectral and angle observations. This problem is known as the decoding problem, and can be solved using the Viterbi algorithm.

4. NUMERICAL RESULTS

In this section, we provide numerical results for the new speaker identification algorithm based on the joint HMM. Results were generated from three test meetings in a medium size conference room of size 18.5' x 9' x 13'. In the conference room, an end wall and a side wall are covered with white-boards creating a highly reverberant environment. Three known participants speak in each of the three experimental meetings; however, the participants vary from meeting to meeting. For the speech spectrum $D = 13$ MDA coefficients are processed for each of the $M = 10$ Gaussian mixtures. Likewise, the angular histogram is composed of $N_a = 45$ bins. The position of the participants for each meeting are shown in figure 2, and the error rates for the three meetings are summarized in tables 2 and 3.

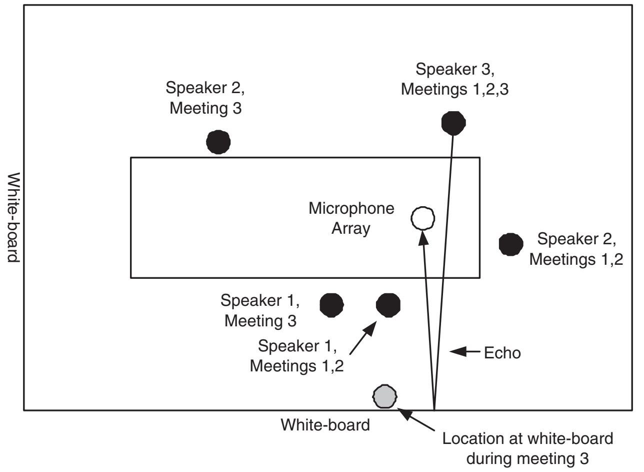

Fig. 2. Room configuration for the three experimental meetings.

In the first meeting, the three participants sat at 0^o, $+90^o$, and -90^o, relative to the placement of the microphone array , and each person spoke for approximately 45 seconds before the next person started speaking. After all three people took turns speaking, the sequence of people speaking was repeated for another 45 seconds each. As shown in figure 2, the microphone array was oriented to produce a strong reflection generated by the third speaker at $+90^o$ due to

Meeting	Speech	Spectrum HMM Error Rate	Joint HMM Error Rate
1	non-overlapping	2.9%	1.8%
2	overlapping	8.2%	5.6%

Table 2. Speaker ID error rates showing significant improvement for the joint HMM algorithm, compared to the spectral only HMM, for the first two experimental meetings with stationary participants.

the white-board. For the meeting, all speech frames were correctly identified except for six. After reviewing the data, all six frames were due to a short burst of laughter. The joint HMM filtered out the laughter and did not switch to the person laughing. By taking the speech spectrum into account, the joint HMM improved the performance compared to an HMM observing the angle alone. Due to the long periods of speech by a single participant, table 2 shows that the joint HMM was able to accurately smooth the results producing an error of 1.8% compared to an error rate of 2.9% based on an HMM using only the speech spectrum.

In the second meeting example, the participants were arranged as in the previous experiment. However, instead of speaking for long durations, a *natural* conversation was recorded with a high number of occurrences of overlapping speech segments between participants. As in the first meeting experiment, the participants remained sitting. In this example, the algorithm also performed very well decreasing the error rate from 8.2% for the speech spectrum only model to 5.6% for the joint model.

The goal of the third meeting was to significantly stress the algorithm having one or more participants sitting far away from the microphone array and having them take turns going to the *same* location at the white-board to write or draw. Again, the participants engaged in natural overlapping conversation and while standing at the white-board, they would often turn toward the white-board while drawing or writing thereby causing a distortion in the spectrum due to the reflected speech. Various 4 minute sections of the recorded meeting were processed at different offsets. Table 3 summarizes the error rates, in percent, at these offsets, in seconds, from the start of the meeting for both the spectrum only HMM and the joint HMM. As the table shows, the algorithm produced a wide variance of error rates with the joint HMM usually providing better results. However, the spectrum only HMM algorithm sometimes performs better (e.g 360 second offset) when there is significant movement as a participant moves to/from the white-board. In this case, the angular information penalizes the spectral estimates. For example with the 360 second offset, speaker 3 was standing at the white-board, returned to his seat, then speaker 1 went to the same location at the white-board and began speaking and drawing. In addition to the moving participants and reflected speech, the signal to noise ratio (SNR) was also fairly low for participants 1 and 2 due to the increased distance from the array shown in figure 2. Participant 2 was located approximately 5.5' away from the microphone array, although still sitting at the conference table. Most likely this low SNR also contributed to the higher error rates compared to the first two meetings where the participants were located closer to the microphone array.

5. CONCLUSIONS

In this paper, we have presented a speaker identification algorithm based on the joint speech spectrum and angle of arrival observations. Including the angle observation significantly improves the error rates for natural meetings with stationary participants as well as filters out spurious vocal noise which is not similar to the speech spectrum. Due to the inherent algorithm smoothing, the joint HMM algorithm does not require additional post processed smoothing that is required for single channel speaker identification algorithms. However, when the participants move significantly and when the SNR is too low, the error rate is too large for practical systems. Additional research is needed to improve the model thereby decreasing the error rates for the case where participants do not remain seated.

Offset	0	120	240	360	480	600	720
Spectrum	41.3	55.2	41.3	45.5	32.6	17.5	23.9
Joint	36.3	67.8	39.6	57.2	18.6	10.6	26.6

Table 3. Speaker ID error rates for the joint and spectral only HMM algorithms for the *extremely challenging*, third meeting with participants moving back and forth to the same location at the white-board and sitting at far distances from the microphone array. The offset times are given in seconds, and the error rates are given in percent for the spectrum only HMM and joint HMM.

6. REFERENCES

[1] D. Moore, "The IDIAP smart meeting room," Tech. Rep. Communication 02-07, IDIAP, 2002.

[2] R. Cutler, "The distributed meetings system," in *Proc. ICASSP 2003*, 2003, vol. 4, pp. IV: 756–759.

[3] J.S. Garofolo, C.D. Laprun, and J.G. Fiscus, "The rich transcription 2004 spring meeting recognition evaluation," in *The Rich Transcription 2004 Spring Meeting Recognition Workshop*, 2004.

[4] J.P. Campbell, "Speaker recognition: a tutorial," in *Proc. of the IEEE*, 1997, vol. 85, pp. 1437 – 1462.

[5] Q. Lin, E. Jan, and J. Flanagan, "Microphone arrays and speaker identification," *IEEE Trans on Speech and Audio Processing*, vol. 2, no. 4, pp. 622–629., 1994.

[6] J. Ortega-Garcia and J. Gonzalez-Rodriguez, "Providing single and multi-channel acoustical robustness to speaker identification systems," in *Proc ICASSP*, 1997, vol. 2, pp. 1107–1110.

[7] I. McCowan, J. Pelecanos, and S. Sridharan, "Robust speaker recognition using microphone arrays," in *Proc. 2001: A Speaker Odyssey*, 2001.

[8] G. Lathoud and I. McCowan, "Location based speaker segmentation," in *Proc. ICASSP*, 2003, pp. III: 621–624.

[9] J. Ajmera, G. Lathoud, and I. McCowan, "Clustering and segmenting speakers and their locations in meetings," in *Proc. ICASSP'04*, 2004, vol. 1, pp. I: 605–608.

[10] J. Bilmes, "A gentle tutorial on the EM algorithm and its application to parameter estimation for gaussian mixture and hidden markov model," Technical Report TR-97-021, ICSI, 1997.

[11] Y. Rui and D. Florencio, "Time delay estimation in the presence of correlated noise and reverberation," in *ICASSP'04*, 2004, vol. 2, pp. ii: 133–136.

[12] H. Malvar, "A modulated complex lapped transform and its applications to audio processing," in *Proc. ICASSP'99*, 1999, pp. 1421–1424.

[13] C.J.C. Burges, J.C. Platt, and S. Jana, "Distortion discriminant analysis for audio fingerprinting," *IEEE Trans. on Speech and Audio Processing*, vol. 11, no. 3, pp. 165 – 174, 2003.