Acoustics'08
Paris
June 29-July 4, 2008
www.acoustics08-paris.org

Speaker identification on the SCOTUS corpus

Jiahong Yuan and Mark Liberman

University of Pennsylvania, 609 Williams Hall, Philadelphia, PA 19104, USA
jiahong@ling.upenn.edu

This paper reports the results of our experiments on speaker identification in the SCOTUS corpus, which includes oral arguments from the Supreme Court of the United States. Our main findings are as follows: 1) a combination of Gaussian mixture models and monophone HMM models attains near-100% text-independent identification accuracy on utterances that are longer than one second; 2) the sampling rate of 11,025 Hz achieves the best performance (higher sampling rates are harmful); and a sampling rate as low as 2000 Hz still achieve more than 90% accuracy; 3) a distance score based on likelihood numbers was used to measure the variability of phones among speakers; we found that the most variable phone is the phone *UH* (as in *good*), and the velar nasal *NG* is more variable than the other two nasal sounds *M* and *N*; 4.) our models achieved "perfect" forced alignment on very long speech segments (40 minutes). These findings and their significance are discussed.

1 Introduction

The U.S. Supreme Court began recording its oral arguments in the early 1950s, and some 9,000 hours of such recording are stored in the National Archives. The transcripts do not identify the speaking turns of individual Justices, referring to them all as "The Court". Therefore, as part of a project to make this material available online in aligned digital form, we have developed techniques for identifying speakers and aligning entire (hour-long) transcripts with the digitized audio.

Information used for speaker recognition may include spectral features [1, 2, 3]; phonetic features [4, 5]; and prosodic features [6, 7]. State-of-the-art speaker recognition algorithms are based on Gaussian Mixture Models (GMM) of spectral measurements, such as Mel Frequency Cepstral Coding (MFCC) features [8] or Perceptual Linear Prediction (PLP) features [9]. The greatest challenge for practical application of speaker recognition is, however, the high variability of channel properties.

In this paper, we demonstrate that GMM-based, monophone HMM models attain robust high performance in speaker identification on the SCOTUS corpus, despite the reverberant environment and the summation of a varying set of microphones. We also establish that our model generates accurate word alignments on hour-long recordings without the need for internal time marks. We suggest that these results not only provide useful information for building robust speaker recognition and word-alignment systems for archival recordings, but they also help develop a better understanding of speaker variability in speech production.

2 Data, Model and Performance

The SCOTUS corpus includes more than 50 years of oral arguments from the Supreme Court of the United States. Seventy-eight arguments from the 2001 term were transcribed, speaker identified, and manually word-aligned by the OYEZ project [10]. Each argument is approximately one hour long. The signal-to-noise ratios (SNR) of the recordings are between 30 and 40 dB. The "clean" turns (based on the transcripts) of eight justices were extracted from these arguments; then, 800 turns (100 for each justice) were randomly set aside as a test set, with the remaining 14,310 turns used for training. The training data is a total of 25.5 hours long. In order to test the robustness of our model to channel variability, we also randomly selected 100 turns (each turn is at least one second long) from the terms of 1995 through 2004 as the second test set.

Our acoustic models are GMM-based, monophone HMMs. Each HMM state has 32 Gaussians Mixture components on 39 PLP coefficients (12 Cepstral coefficients plus energy, and Delta and Acceleration). First, we trained a general acoustic model using all the training data; then, we adapted it to justice-specific models with each justice's data. Our language models were justice-specific, phone-based bigrams. Both the acoustic and language models were trained using the HTK toolkit [11] and the CMU Pronouncing Dictionary [12]. Normal speech recognition was conducted eight times for each utterance, using the models of the eight justices, with the highest-scoring model used to identify the speaker. The dictionary used for recognition contained only monophones; each monophone has a pronunciation of itself. Again, the HTK toolkit was used for the decoding.

Our models achieved 98.0% accuracy on the 800 randomly selected test segments from the 2001 term. The 16 errors (2.0%) are all "false errors"; either these segments contain significant overlaps between different speakers or high background noise, or they are too short (much less than one second). The test on the 100 turns from the 1995 to 2004 terms also showed perfect results, although this test data used different recording devices and was digititized at different sampling rates. Three of the 100 turns were not correctly identified, but all of them are "false errors".

3 Effects of Sampling Rates

Our training data and first test set were extracted from the 2001 term and were sampled at 44,100 Hz; however, the arguments from earlier terms were sampled at either 44,100 Hz or 22,050 Hz. We therefore downsampled the data for both training and test. Table 1 shows the accuracies of the models that were trained on the same data set but with different sampling rates. The first test set (800 turns from the 2001 term) was used for testing. The test turns were downsampled to the same sampling rate as the training data.

The results in Table 1 suggest that the sampling rate of 11,025 Hz has the best performance. Surprisingly, a sampling rate as low as 2,000 Hz can achieve more than 90% accuracy. This result suggests that most of the inter-speaker variability is conveyed in frequencies below 1,000 Hz in the speech signal, at least in these recordings.

Sampling rate	Accuracy
2000 Hz (39 parameters)	92.9%
4000 Hz (39 parameters)	96.4%
8000 Hz (39 parameters)	97.6%
11025 Hz (39 parameters)	**98.0%**
22050 Hz (39 parameters)	96.4%
44100 Hz (39 parameters)	94.8%
44100 Hz (60 parameters)	96.8%

Table 1. Effects of sampling rates.

Except for the final model, all the models in Table 1 used 39 PLP coefficients, and the features were extracted using 12 filters spread over the frequency range from zero to the Nyquist frequency. The last model used 60 PLP coefficients (19 Cepstral coefficients plus energy, and Delta and Acceleration) and 32 filters from zero to the Nyquist frequency of 22,050 Hz; it used almost the same mel-scale filters from zero to 11,025/2 Hz as the 11,025 Hz model, plus more filters on the higher frequencies. Table 1 shows that this model is still not as good as the 11,025 Hz model; these results suggest that there is no useful speaker information in frequencies above 5,500 Hz in these recordings, at least for the approach we used.

4 Inter-Speaker Variability

Since the GMM-based, monophone HMM models perfectly capture inter-speaker variability, we decided to investigate how and where the speakers differ, as seen by the models. We had two goals: to help design the prompt text for text-dependent speaker identification systems (the phones that are more distinctive among the speakers should be used); and to learn more about speaker variation and phonetic variation in general.

To study which phones are more distinguishable among the justices, we defined a distance score, *D*, based on the likelihood numbers from different speaker models. In the equation (Eq.(1)), D(p$_i$, s$_j$) is the distance score of the phone-token p$_i$ made by the speaker s$_j$; L(p$_i$, M$_j$) is the likelihood score of p$_i$ when it is forced aligned using the model M$_j$, the speaker s$_j$'s model; and *N* is the number of speakers.

$$D(p_i, s_j) = \frac{\sum_{k=1,2,\ldots N; k \neq j} \left| L(p_i, M_j) - L(p_i, M_k) \right|}{N-1} \qquad (1)$$

We then calculated the Mean Distance Score for each phone type of each speaker. Mean(D(/a/, "*Tom*")), for example, is obtained by averaging the distance scores of all the tokens /a/ from Tom; it measures the 'deviation' of Tom's /a/ from the other speakers' /a/. Figure 1 below shows the average Mean Distance Scores of the eight justices for each phone; these scores were calculated based on the 800 utterances in the first test set. The higher the distance score, the more distinguishable the phone is among the eight justices.

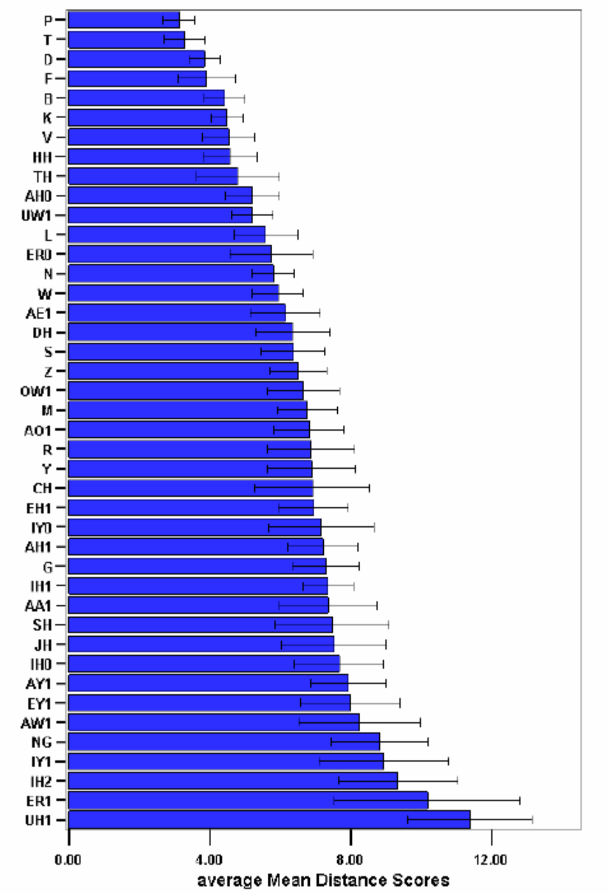

Fig. 1. Phone variability measured by Mean Distance Score.

Figure 1 shows that vowels carry more speaker variability than consonants; that the the most variable phone is the phone *UH* (as in the word *good*); and that the velar nasal *NG* is more variable than the other two nasal sounds *M* and *N*. Further studies are needed to explain these results.

5 Forced Alignment

Another major finding of this study is that accurate word alignment on long speech segments can be achieved by using the HTK toolkit and our acoustic models. Figures 2 and 3 show the forced alignment errors on speech segments that are between 10 to 40 minutes in length. Figure 2 shows the difference between the forced aligned word boundaries and the hand-labeled word boundaries; the differences are consistently around 50 milliseconds along the time of word onset. Figure 3 displays the histogram of the differences.

Since the speech segments contain pauses, noises, disfluencies, etc., and because the hand-labeled boundaries are not 100% accurate, the "true errors" of the forced alignment are shorter than 50 milliseconds. In fact, we found that most of the forced-aligned word boundaries are "perfect".

There are several possible reasons that our acoustic aligner works better than some others have in the past: 1) the training data is large and fairly clean; 2) we used GMM-based, monophone models instead of triphone models; and 3) we corrected a rounding issue that arises when using the HTK toolkit to extract features. If the sampling rate is 11,025 Hz and the time step is set to 10 milliseconds, then the analysis window will move forward by 110 samples instead of 110.25 samples at each step. If the speech segment is one hour long, the time difference between the original speech signal and the extracted feature vectors can be as long as eight seconds (0.25 samples x 360,000 = 90,000 samples). We adjusted this time difference when conducting forced alignment.

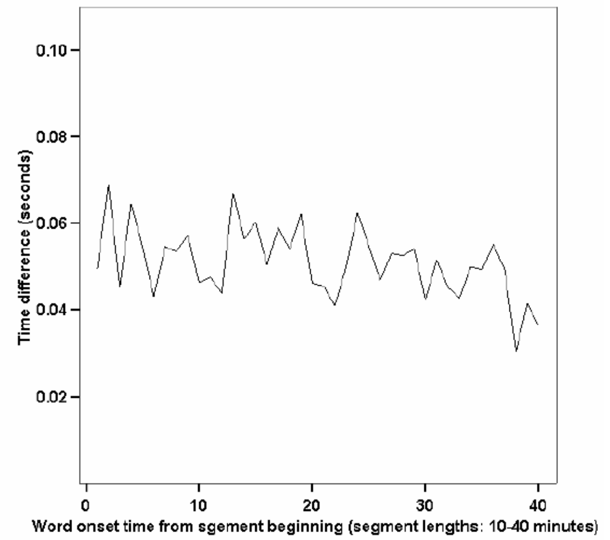

Fig. 2. Forced alignment errors at different word onset time.

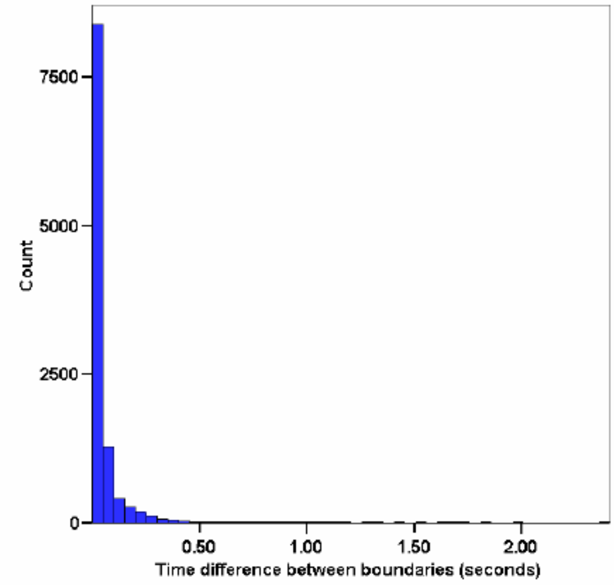

Fig. 3. Histogram of forced alignment errors.

6 Conclusions

GMM-based monophone HMM models can generate excellent results in justice identification on the SCOTUS corpus. Most of the inter-speaker variability is conveyed in frequencies below 1,000 Hz in the speech signal; and modeling features extracted from the frequencies above 5,500 Hz are generally harmful to speaker identification. Vowels carry more speaker variability than consonants; the most variable phone is the phone *UH* (as in the word *good*); and the velar nasal *NG* is more variable than the other two nasal sounds *M* and *N*. Not only do these results provide useful information for building robust speaker recognition systems, but they also help develop a better understanding of speaker variability in speech production. Our study also demonstrated accurate word alignment on very long speech materials, which is significant for many practical applications.

Acknowledgments

We would like to thank Jerry Goldman and the team of the OYEZ project for providing the data. This work was partially supported by NSF award 0325739.

References

[1] J.P. Campbell, “Speaker Recognition: A Tutorial,” *Proceedings of the IEEE*, vol. 85, no. 9, pp. 1437-1462, 1997.

[2] D.A. Reynolds and R.C. Rose, “Robust text-independent speaker identification using Gaussian mixture speaker models,” *IEEE Trans. Speech Audio Processing*, vol. 3, no. 1, pp. 72-83, 1995.

[3] C.S. Liu, C.H. Lee, W. Chou, B.H. Juang, A.E. Rosenberg, “A study on minimum error discriminative training for speaker recognition,” *J. Acoust. Soc. Am.*, vol. 97, pp. 637-648, 1995.

[4] F. Nolan, *The Phonetic Bases of Speaker Recognition,* Cambridge: CUP, 1983.

[5] J.A. Bachorowski and M.J. Owren, “Acoustic correlates of talker sex and individual talker identity are present in a short vowel segment produced in running speech,” *J. Acoust. Soc. Am.*, vol. 106, no. 2, pp. 1054-1063, 1999.

[6] E. Shriberg, L. Ferrer, S. Kajarekar, A. Venkataraman, and A. Stolcke, “Modeling prosodic feature sequences for speaker recognition,” *Speech Communication*, vol. 46, no. 3-4, pp. 455–472, 2005.

[7] A.G. Adami, “Modeling prosodic differences for speaker recognition,” *Speech Communication*, vol. 49, no. 4, pp. 277-291, 2007.

[8] S.B. Davis and P. Mermelstein, “Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences,” *IEEE Trans. ASSP*, vol. 28, no. 4, 357-366, 1980.

[9] H. Hermansky, “Perceptual linear predictive (PLP) analysis of speech,” *J. Acoust. Soc. Am.*, vol. 87, no. 4, pp. 1738-1752, 1990.

[10] http://www.oyez.org/

[11] http://htk.eng.cam.ac.uk/

[12] http://www.speech.cs.cmu.edu/cgi-bin/cmudict/