INTERSPEECH 2009 BRIGHTON

The Multi-Session Audio Research Project (MARP) Corpus: Goals, Design and Initial Findings

A. D. Lawson[1], A. R. Stauffer[1], E. J. Cupples[1], S. J. Wenndt[2], W. P. Bray[3], J. J. Grieco[2]

[1]RADC, Inc., Rome, NY USA
[2]Air Force Research Laboratory, Rome, NY USA
[3]Oasis Systems, Lexington, MA USA

Aaron.Lawson.ctr@rl.af.mil, stauffar@clarkson.edu, Edward.Cupples.ctr@rl.af.mil, Stanley.Wenndt@rl.af.mil, Wayne.Bray.ctr@rl.af.mil, John.Grieco@rl.af.mil

10.21437/Interspeech.2009-528

Abstract

This project describes the composition and goals of the Multi-session Audio Research Project (MARP) corpus and some initial experimental findings. The MARP corpus is a three year longitudinal collect of 21 sessions and more than 60 participants. This study was undertaken to test the impact of various factors on speaker recognition, such as inter-session variability, intonation, aging, whispering and text dependency. Initial results demonstrate the impact of sentence intonation, whispering, text dependency and cross session tests. These results highlight the sensitivity of speaker recognition to vocal, environmental and phonetic conditions that are commonly encountered but rarely explored or tested.

Index Terms: corpus development, speaker identification

1. Introduction

1.1. Goals of the current paper

This paper provides an introduction to the Multi-session Audio Research Project (MARP) corpus, detailing the motivation for the corpus of the project, the research questions the corpus was designed to help clarify and some preliminary findings. This is important because while the MARP corpus has begun to be exploited for work in session variability, speaker variability mitigation [1] and exploring the impact of data conditions on speaker recognition (SR) [2] there has been no formal citation for this corpus. Furthermore, the very interesting and eye-opening results obtained by testing the specific conditions for which the corpus was designed have not been made available to the speaker recognition community in a formal publication. It is believed that the factors tested can provide clues to the improvement of SR algorithms and directions for mitigation of condition mismatch.

1.2. Background

Speaker Recognition is notoriously impacted by mismatch between training and testing conditions. Differences in speech conditions such as noise [3], channel [4], vocal register (Lombard speech [5], etc.) are widely acknowledged to play a negative role in SR. On the other hand, text dependency between test and train data is widely believed to improve SR performance. There are many additional factors relating to the conditions under which SR is performed that are less well understood than channel, noise, etc. but which may have a significant impact on SR accuracy. A second goal of this study is to determine the effect of some of these factors, including whispered speech vs. phonated speech and the impact of intonation, detailed in the next section.

2. The Multi-Session Audio Research Project (MARP) Corpus

2.1. Goals of the database

The purpose of the MARP database was to provide audio recordings to allow for the testing of six speaker identification parameters, and their effects on speaker identification accuracy: 1) the effect of time or aging, 2) inter-session variability over a great number of sessions, 3) the impact of the speaker's intonation, 4) whispered speech, 5) text dependency over time, and 6) the difference between read and spontaneous speech. To accommodate interest in the effects of time, aging, and intersession variability the MARP Corpus was designed to consist of multiple sessions of the same speakers recorded as 21 sessions over a three-year period of time.

2.2. Components of the database

There are five major components that make up each recording in the MARP Corpus; read spoken sentences, whispered sentences, a read passage, read digits, and a conversation. The first section is a series of read spoken sentences designed phonetically to be maximally diverse. Within this section are ten declarative sentences, followed by the same ten sentences spoken with different intonations: four spoken as exclamations, and the remaining six are repeated in the form of a question. These read sentences remained constant throughout the entirety of the MARP collect, to allow for the testing of the effect of text dependency on speaker ID. Ten extended-length sentences are also included in this section, half of which remain the same throughout all sessions, while the other five are chosen at random from a fixed set each time.

The section of whispered speech consists of ten read, whispered sentences. The first five were chosen from the first set of spoken sentences and the rest were taken from the extended spoken sentences. Sentences used for the whispered speech section remained the same throughout the ten sessions where whispered speech was collected. Participants were trained in what constituted whispering and all whispering was 100% monitored to ensure that speakers actually whispered.

For each session, a one to two minute passage is randomly chosen to be read by every speaker. These passages are unique to each session, and are read straight from a prompt. This same procedure is followed to choose the random digit strings that follow the read passage.

The final component of each MARP collect is a free-flowing conversation of approximately ten minutes in length, with an isolated partner. Speakers were encouraged to keep

participation equal on both sides of the conversation. Speakers were also given suggested conversation topics and a monitor ensured that the dialogue filled the entire allotted time. Session partners remained constant throughout the three years. A total of 1,080 conversational sides were collected.

All sessions were recorded in an anechoic chamber using high quality microphones. In total 540 sessions were completed. Audio was recorded at 24 bit resolution and sampled at 48000 Hz in pcm format, digitized using a Edirol FA101 firewire capture. For SR experiments audio was converted to 16 bit, 8000 Hz pcm.

3. The Speaker ID System

The Gaussian Mixture Model (GMM) and Universal Background Model (UBM) approach, developed by Reynolds [6], is used as the basis for speaker recognition in this study. In our implementation the front-end feature processing consists of mel-weighted and delta cepstra generated from a frame size of 20ms with 50% overlap. During recognition, the likelihood of the test speech is computed for each of the GMMs produced during training. Only 5 mixtures are used for the calculation of the likelihood of a particular speaker's GMM model, and the five mixtures are chosen from the most probable mixtures in the UBM. This study does not focus specifically on the accuracy of a given speaker recognition system in order to compare or improve algorithms, rather the goal is to demonstrate the effect of the experimental conditions on a very common approach to speaker recognition.

4. Experiments

Each experiment involved 37 speakers who occurred in all the sessions tested. The GMM-UBM system was trained with 128 mixtures because this configuration performed optimally in calibration experiments, probably due to the small size of the training and test data. Each model was trained on approximately 8 seconds of audio and adapted to a UBM built from a similarly-recorded database, TIMIT. Cepstral Means Subtraction and silence removal were not employed because it was found to decrease accuracy in these experiments. All tests were scored using a forced decision/closed set approach where the #1 top ranked model must match the true speaker to be considered correct, all other results were marked as error.

5. Cross-Session Experimental Setup

The first condition examined whether speaker recognition accuracy was affected by testing across sessions. Data consisted of five separate sessions spread over a year: months 3, 7, 8, 9 and 12. Test and train data consisted of phonetically non-identical sentences, i.e. text independent. Models for each speaker were trained on six of the "short sentences", approximately 8 seconds of speech in total. This was tested on four sentences, each averaging about 1.4 seconds in length. Every session was tested against every other session and results were obtained by comparing cross-session results to in-session results.

5.1. Intonation Experimental Setup

All intonation tests where performed in-session. Intonation data consisted of interrogative and exclamatory sentences. To avoid expected participant error in producing these utterances each subject was trained to understand what intonation was expected. Each sentence was also accompanied by an auditory prompt of the sentence spoken with the target intonation. Training data was approx. 8 seconds/6 sentences per model. Test data consisted of 4 sentences per speaker, two interrogative and two exclamatory for intonation testing and four declarative sentences or the reverse condition.

5.2. Whispered Speech Experimental Setup

Pilot training showed that some participants were not aware of how to whisper. To counter this all participants were thoroughly trained as to what was meant by "whispering" and an observer listened in on all recordings to insure that subjects were whispering. When whispered data was not produced subjects were asked to repeat the sentence. Each session produced 6 whispered sentences. Test and train data was all in-session, with no text dependency between sets. Two conditions were tested: phonated models on whispered test and whispered models on phonated test. Whispered on whispered could not be tested at this time due to insufficient in-session, non-text dependent data. This has been remedied by increasing the number of whispered sentences in later sessions.

5.3. Text Dependency

Training data consisted of four sentences each (2, 3, 7, 9) from four sessions. Test data consisted of sentence 2 (from sessions not included in train) plus three other sentences that were phonetically distinct from the training data. In this part of the study two questions were addressed. The first is rather simple: How much does text dependency (phonetic identity between test and train) improve SID performance, if at all? This set up is depicted in figure 1:

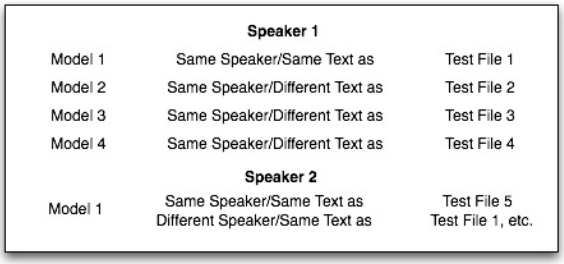


Figure 1: *Text Dependent Test Set up*

The second is more complex: to determine whether a speaker recognition system could be tripped up by matching phonetic content.

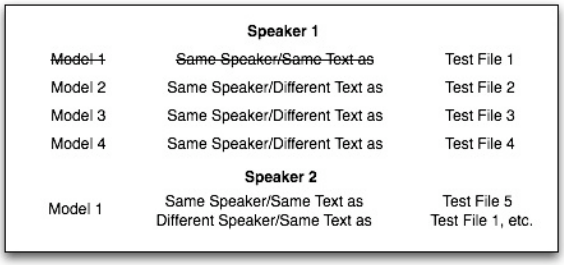


Figure 2: *Text Dependent Model Removed for Current Target Only*

This set up verifies if the SR system identifies the wrong speaker uttering a matching sentence over the correct speaker saying a different sentence; i.e. do speaker characteristics

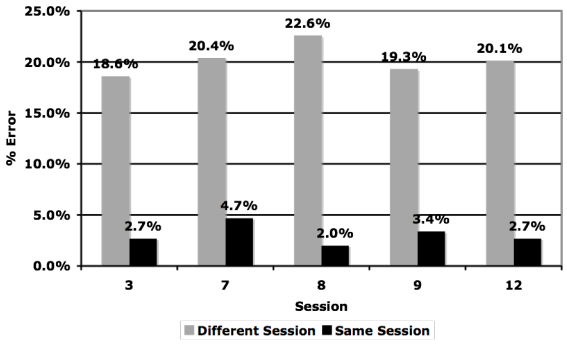
trump phonetic content, or vice versa? For this scenario, the text dependent model (made from sentence 2) is removed for the target speaker, but left in for all other speakers, to determine if the top scoring model would come from the target speaker (ignoring textual identity for speaker identity, or if another speaker's model would win (choosing textual identity over speaker characteristics). This setup is depicted in figure 2.

6. Results

6.1. Cross-Session Impact

Cross-session tests revealed a very strong trend. As one can see in table 1, in-session tests averaged 3.11% error, cross-session tests averaged 20.2% error, a total difference of 17% more error in cross-session over same-session tests.

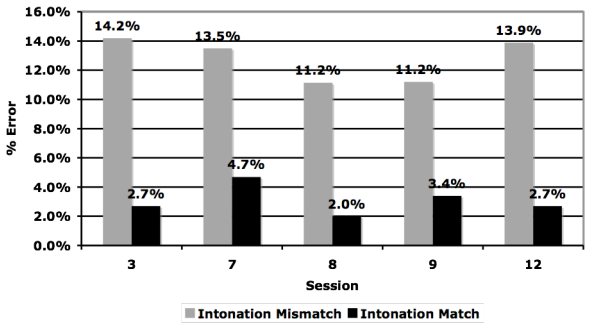
Table 1. *Same session vs. different session speaker recognition error rates*



6.2. Impact of Intonation

Cross-intonation tests showed an average increase in error of 9.7% over the baseline matched condition tests. Exclamatory/interrogative intonation train on declarative test had an average error of 10.1%. The Declarative intonation train set on exclamatory/interrogative intonation test had higher error, averaging 12.8%. As one can see in table 2 the impact of intonation was common across all sessions tested.

Table 2. *Impact of matched intonation vs. mismatched intonation on speaker recognition*



6.3. Impact of Whispering vs. Normal Phonation

In table 3 one can see that whispering has a dramatic impact on speaker recognition performance, with an average error rate of 84.3%. This effect was as strong whether one trained on whispered speech and tested on phonated or vice versa.

Table 3. *Whispered speech*

	Error Rate
Whispered Train on Phonated	82.3%
Phonated Train on Whispered	86.3%
Average	84.3%

6.4. Impact of Text Dependency

In "text-dependent" tests there was 0% error, despite the fact that these tests were all run in cross-session set-up, which resulted in over 20% error with text independent data (see 5.1). In every single case the model chosen matched the text of the test clip, even though there were three other models for that speaker.

Table 4. *Text dependency*

Matched Text Dependency (% Error)	Top Model Sentence = Test Sentence	Matched Sentence Removed (% Error)	Top Model Sentence = Test Sentence
0%	100%	91.9%	91.9%

When the "reverse text dependent" scenario (all other speakers had models whose text matched the test clip text except the current target speaker) was tested, error jumped to 91.9%, as can be seen in table 4. Every time the wrong speaker model won that model's training text matched the test clip's text.

7. Discussions/Conclusions

7.1. Cross-session

Even in an anechoic chamber with identical microphones and consistent conditions inter-session results show much higher error than same session. It is unclear why this condition had such an impact, since the usual compromising conditions impacting SR, channel, noise, signal degradation, etc. were not in play. This can only be the result of differences in the speakers' voices, a natural "micro register" variation. Seasonal effects were tested and they did not play a role. These results spurred significant follow up research into the nature of inter-session variability with the goal of identifying its correlates for mitigation [1]. Clearly if the root causes of the discrepancy can be identified and accounted for SR's reliability across time can be greatly improved.

7.2. Intonation

A second conclusion was that differences in intonation patterns do have a significant effect on SR. This points to role of alterations in the formant structure caused by changes in F0 as potential factors in SID accuracy, as both intonations in English tend to raise F0 and change prosody. Sentences spoken with exclamatory intonation tend to have greater energy due to their emphatic nature. As one can see in figure 3, as pitch changes the resonance bands and formants of the vowel change throughout the spectrum. An increase in F0 induces a corresponding increase in the distance between resonance bands, a rise in frequency of F2 and F3 and a reduction in the bandwidth of the resonance bands and an increase in the bandwidth of the formants.

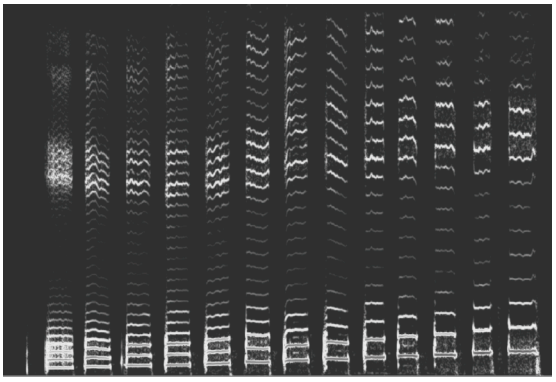


Figure 3: *Impact of male subject increasing pitch from 85Hz to 255Hz while saying the phoneme /i/*

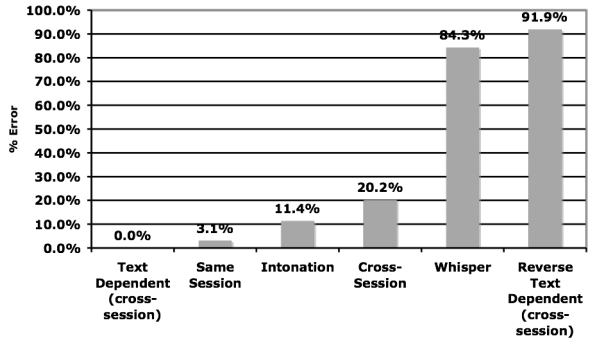
7.3. Whispering vs. Phonation

The impact of whispered speech was dramatic, but not so surprising. These findings certainly help reveal what characteristics of the audio signal are necessary for speaker recognition to perform accurately. It may be the case that phonation is critical to speaker identification in an MFCC system, to verify this experiments with whispered training on whispered test are in order.

7.4. Textual/Phonetic Content

Finally, there is clearly a complex relationship between phonetic content and speaker identification. Text dependency improved results enormously across sessions, yet it is clear that SR systems can be severely mislead by textual content. As one can see in table 5 the greatest source of error came from the "reverse text dependence" test, which showed that given a preference between matching phonetic content and speaker information, the phonetic content wins out.

Table 5. *Average Error Rates for Conditions Evaluated*



This begs the question: how many SR errors are caused by phonetic content? In one sense, for SR purposes the phonetic content may be seen as a kind of noise: in matched conditions it is normalized out, but under other circumstances it can force a wrong decision because the phonetic match is clearly influencing the outcome. Since the primary function of the speech signal is to carry phonetic information, and only secondarily speaker information, separating speaker characteristics from content poses a significant problem. This finding reinforces the importance of phonetic content for speaker verification purposes and corroborates earlier findings [7] that phonetic overlap (measured in tri-phones) between test and train data correlates highly (> .97) with SID performance.

7.5. Conclusions/Observations

It is safe to say that classifiers do an inordinate amount of work in SR because SID features do not appear to capture speaker characteristics very well. This may be unavoidable, since the primary function of speech is transmission of phonetic content, not speaker characteristics. Better SR features would identify factors that distinguish speakers and minimize phonetic content. This might not be possible with the current fixed frame signal decomposition; longer time-frame, non-modal, features may prove more accurate under the conditions examined in this study. Indeed, adaptations that take direct advantage of voice characteristics at a low level, such as [1] have been shown to provide improvements in SR across sessions using standard MFCCs on the MARP corpus.

7.6. Future Work

The MARP corpus is being prepared for distribution through the Linguistic Data Consortium, and will hopefully be available in late 2009. A follow up to the MARP corpus collection of data is beginning and will include several additional factors: 1) five microphones in different locations in the anechoic chamber with different frequency responses and bandwidths, 2) introduction of varying degrees of natural noise to the speakers' headphones, but not to the recordings themselves, to measure the impact on speaker recognition of the pure vocal effect induced by communication in different environments. This same noise can be added to the recordings to separate the effect of noise contamination on SID from the alterations in the voice noise induces.

8. References

[1] Lawson, A., Linderman, M., Leonard, M., Stauffer, A., Pokines, B., Carlin, M. "Perturbation and pitch normalization as enhancements to speaker recognition" ICASSP 2009, Taipei, Taiwan.

[2] Lawson, A., Stauffer, A, Wenndt, S., "External factors impacting the performance of speaker identification in the Multisession Audio Research Project (MARP) corpus", *153rd Meeting of the Acoustical Society of America*, June 4-8, 2007.

[3] J. Ming, T. Hazen, and J. Glass, "A Comparative Study of Methods for Handheld Speaker Verification in Realistic Noisy Conditions," Proc. of IEEE Odyssey 2006: The Speaker and Language Recognition Workshop, 2006.

[4] Reynolds, D. A. "Channel robust speaker verification via feature mapping", ICASSP 2003, Hong Kong, China.

[5] Ikeno, I. and Hansen, J. "Perceptual in-set speaker identification using neutral speech and speech under Lombard effect", *Proceedings of the International Association for Forensic Phonetics and Acoustics*, Göteburg, Sweden, 2006.

[6] Reynolds, D. A. "Comparison of Background Normalization Methods for Text-Independent Speaker Verification." *Proceedings of the European Conference on Speech Communication and Technology*, Rhodes, Greece, 1997. Vol. 2, pp. 963-966.

[7] Lawson, A. and Huggins, M. "Triphone-Based Confidence System for Speaker Identification", *Interspeech 2004*, Jeju, Korea.