Session Variability Contrasts in the MARP Corpus

Keith W. Godin, John H.L. Hansen

Center for Robust Speech Systems
Univ. of Texas at Dallas, Richardson, TX, USA
godin@ieee.org, john.hansen@utdallas.edu

Abstract

Intra-session and inter-session variability in the Multi-session Audio Research Project (MARP) corpus are contrasted in two experiments that exploit the long-term nature of the corpus. In the first experiment, Gaussian Mixture Models (GMMs) model 30-second session chunks, clustering chunks using the Kullback-Leibler (KL) divergence. Cross-session relationships are found to dominate the clusters. Secondly, session detection with 3 variations in training subsets is performed. Results showed that small changes in long-term characteristics are observed throughout the sessions. These results enhance understanding of the relationship between long-term and short-term variability in speech and will find application in speaker and speech recognition systems.

Index Terms: speaker identification, session variability

1. Introduction

Speaker intersession variability is a known problem for speech systems, including speaker identification systems. When testing on data from a session not represented in models, a drop in identification performance is to be expected compared to testing on data from a session represented in speech models. Performance problems due to intersession variability are often related to changes in microphone or channel, but recent work has focused on identifying those aspects of the performance problem that are due solely to changes in the speech production behavior of the speaker. Such work is supported, for example, by the Multisession Audio Research Project (MARP) corpus [1]. The MARP corpus, introduced more fully below, includes many sessions over a long period of time (3 years) from several speakers, all recorded in the same facility using the same equipment, supporting isolated studies of changes in speech production.

In this study, intrasession variability is defined as speech production traits that are observed to change in the course of one session, and intersession variability as speech production traits that are observed to change between sessions but that stay constant within one session. These working definitions have only vague association with phenomena of speech production, but form a useful framework for studies in this area. This framework is useful because it relates directly to the framework employed for speaker identification. The analysis of such systems is concerned with the ways in which speech production might vary within the timeframe of a typical session, and between such sessions, especially because systems often assume the availability of speech training data from only one session.

This project was funded by AFRL through a subcontract to RADC Inc. under FA8750-09-C-0067, and partially by the University of Texas at Dallas from the Distinguished University Chair in Telecommunications Engineering held by J. Hansen. Approved for public release; distribution unlimited.

A recent study employing the MARP corpus [2] showed several interesting results regarding long term speech variability, two of which relate to and motivate the work in this study. The first result was that there is a relationship between 30-second chunks of conversational speech taken from the beginning of different sessions, which was found by training models of early 30-second chunks from one session and comparing test scores of that model against chunks from other sessions. The present study seeks to determine the extent to which relationships between 30-second chunks are intersession or intrasession, to determine whether unseen intrasession variability may be the cause of the observed intersession performance problems. This is the focus of the first experiment, which uses a distance measure to cluster models of 30-second chunks of speech.

The second result from [2] that is related to the present study was that aging is not a factor in intersession performance problems, as only a weak relationship was observed between model scores and time between training session and test session. Thus, medium-term speech variability such as fatigue, circadian rhythm, and mood are more likely the cause of observed performance problems. This study seeks to confirm and expand on that result, determining more generally whether sessions are related to neighbors in time as closely as to those recorded more distantly in time. This is the focus of the second experiment of this study, which explores session verification using three different training regimes.

2. Corpus

The Multisession Audio Research Project (MARP) corpus [1] is employed in this study for the study of intrasession and intersession variability. The MARP corpus offers a unique opportunity to compare intersession with intrasession variability. It contains speech over a period of three years from a number of subjects, recorded at intervals of 3-6 months. Speech recorded for each session includes a 10-minute conversation on suggested but varying topics, and a read portion that includes a variety of speaking styles, including question and whisper. For this study, only the conversational portions of each session are employed. This study employs 14 speakers, each with 17-19 10-minute sessions of conversational speech.

3. Experiment 1: Clustering 30s chunks

The experiment consists of using a distance measure to build clusters out of 30-second chunks of the sessions, clusters which should in theory be made up of chunks that are similar to each other in the strongest ways that the modeling technique employed may detect. By examining the intrasession vs. intersession composition of the clusters, research questions about

the relative importance of intersession variability, as compared to intrasession variability, may be addressed. Addressing these questions may offer new evidence in support of theories about the causes of the observed performance drop in speech systems due to intersession variability.

The experimental method employed in this study comprises 5 steps. The results of the following experiment are analyzed in Section 4:

1. Break each session into 30s chunks, resulting in 20 chunks per session
2. Model each chunk of speech with a Gaussian Mixture Model (GMM)
3. Determine the two most closely related chunks (clusters)
4. Collapse the two closest chunks into one cluster, and create a new model for that cluster
5. Repeat steps 3 and 4 until just 20 clusters remain

The distance measure employed in this experiment is the symmetric Kullback-Leibler divergence. The GMMs are trained using the Hidden Markov Model Toolkit (HTK) [3]. The KL divergence code were written in-house. The symmetric Kullback-Leibler (KL) divergence [4] is a method for measuring the relatedness of two probability density functions. It is not a metric, as it is not 0 for 'equal' objects, but is a popular and robust measure of the similarity of two GMMs.

4. Experiment 1: results analysis

The analysis of the experimental results addresses the following questions:

- What is the degree of cross-session clustering?
- Is there any consistent trend to clustering beginning, middle, and end of sessions?
- Is there a trend to cluster contiguous blocks of speech?

4.1. Cross-session clustering

The degree of cross-session clustering is investigated by determining the average number of sessions clusters are drawn from. Given, for example, a speaker from whom 18 sessions were recorded, it may be said that if most clusters are drawn from fewer than 5 sessions, intersession variability dominates, causing 30-second chunks of speech to be more closely related to other chunks from the same session. If the average number of sessions clusters are drawn from is around 14 or 15, it may be said that intrasession variability dominates.

It is also important to determine whether the clustering results are meaningful, or whether, due to unforeseen problems with the clustering method, merely random. To determine this, the frequency of occurrence of the number of sessions clusters are drawn from is compared to a theoretical probability mass function that would apply if the clustering were random. It is straightforward to apply the Kolmogorov-Smirnov (KS) test here to compare the empirical frequency of occurrence with a theoretical probability mass function, as long as a suitable PMF may be derived.

4.1.1. Probability mass function for random clustering

In this section is derived the probability mass function (PMF) describing the number of sessions clusters would be drawn from if the clustering were random. If a cluster has k chunks, a speaker was recorded in s sessions, and a cluster is formed using chunks from exactly n sessions, then the number of ways to form such a cluster is

$$c(s,k,n) = \binom{s}{n}\binom{k-1}{n-1}. \quad (1)$$

The simplifying approximation is made that clusters are all of the same size k and furthermore that $k = s$. This may be made because, if clustering is random, the expected value of the cluster size is equal to the total number of chunks available ($s * 20$), divided by the number of clusters. For each speaker, the analysis is always performed on a clustering result where s clusters were formed, thus, if the results are random, on average $k = s$.

Finally, to form a probability mass function, $c(s = k, n)$ is evaluated for each n for each of the three values $s = k$ available in the corpus (17 sessions, 18 sessions, and 19 sessions), and each set of results over all possible values of n is normalized to 1 across s. For each n, the final probability mass function is formed by weighting each component PMF by the relative number of speakers who spoke over that many sessions (7 speakers had 17 sessions, 5 speakers had 18 sessions, and 2 speakers had 19 sessions). Thus, if $t(s)$ speakers were recorded in s sessions (i.e. $t(17) = 7$), and there are 14 speakers,

$$p(n) = \sum_{s=17}^{19} \frac{t(s)}{14} \frac{\binom{s}{n}\binom{s-1}{n-1}}{\sum_{i=1}^{s}\binom{s}{i}\binom{s-1}{i-1}} \quad (2)$$

The resulting PMF is shown in Fig. 1.

4.1.2. Cross-session clustering results

Fig. 1 shows both the measured frequency of occurrence of clusters formed from chunks from the given number of sessions, and the theoretically predicted PMF if the clustering is random. A Kolmogorov-Smirnov (KS) test, performed with an alpha of $\alpha = 0.01$, shows that the measured frequency of occurrence is different from the predicted probability mass function, which is consistent with inspection of the figure. The mean of the empirical results is 12.9. Therefore, it is concluded both that the results of the clustering are not random, and that, when modeled with an MFCC-GMM technique, intrasession variability between 30-second chunks of speech dominates over intersession variability.

Min. average purity	0.5017
Max. average purity	0.5312
Mean average purity	0.5157

Table 1: *Statistics of average purity level for each speaker*

4.2. Trends to clustering conversational parts

The analysis of this section investigates whether there is a trend to form clusters from conversational parts represented by the beginning, middle, and end of sessions. The conversational "purity" of the clusters is here defined as the maximum of their percentage makeup of the beginning (first 5 chunks), middle (middle 10 chunks), or end (last 5 chunks) of sessions, and forms a rough, though practical, approximation of conversational parts for analysis. Random clustering, as well as clustering unrelated to these divisions, will result in purity levels close to 50%, as 50% of possible chunks are drawn from the middle sections. If clusters generally have high purity, it may be that conversational

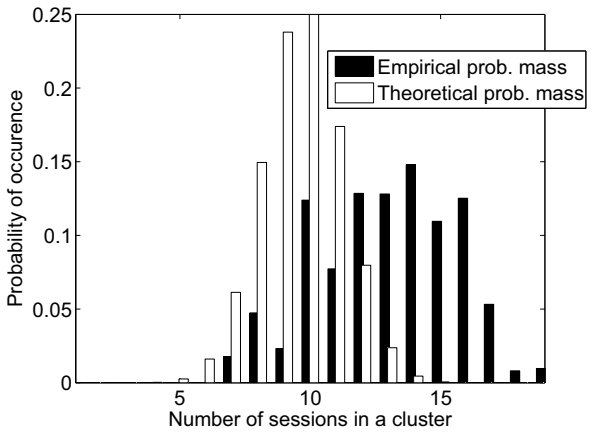

Figure 1: *A comparison of the empirically derived probability mass function of the number of sessions in a cluster, with a theoretical prediction based on the assumption the clustering is random.*

Min. cluster size	7
Max. cluster size	46
Std. dev. cluster size	7.63

Table 2: *Statistics of cluster size over all speakers. The mean is by definition 20.*

tone, or this approximate measure thereof, is a measurable form of intrasession variability.

For each speaker, the average purity level of clusters of chunks of his or her speech is measured. The minimum, maximum, and mean of these averages is shown in Table 1. Statistical tests for a mean different from 50% are precluded by the number of speakers (14), but because the mean is greater than 50%, and none of the 14 averages lie below 50%, the three statistics shown make a preliminary case for conversational tone as a form of intrasession variability captured by this study's modeling technique.

4.3. Cluster size

Finally, a brief look at the size of the clusters is warranted. Because the chunks from each speaker are clustered into the same number of clusters as sessions available for that speaker, the mean cluster size is by definition 20 chunks (the size of one session). Table 2 shows the minimum, maximum, and standard deviation of cluster sizes across all speakers. Subjectively, there is a fair amount of variance among cluster sizes, indicative of nonrandom clustering.

5. Experiment 2: Session verification

Having learned from the first experiment that intrasession variability dominates the relationships between 30-second chunks of speech, this second experiment explores whether performance problems due to intersession variability in other studies are in fact due to unseen intrasession variability. For this purpose is adopted a session verification experiment paradigm. Session verification is an adaptation of the speaker verification paradigm to differentiating recorded sessions of one speaker. In the following experiment, the performance of session verification is evaluated for each speaker, and conclusions are drawn from the average performance across speakers. While session verification itself does not have practical application, inferences may be made from this experiment about the relationship between intersession and intrasession variability.

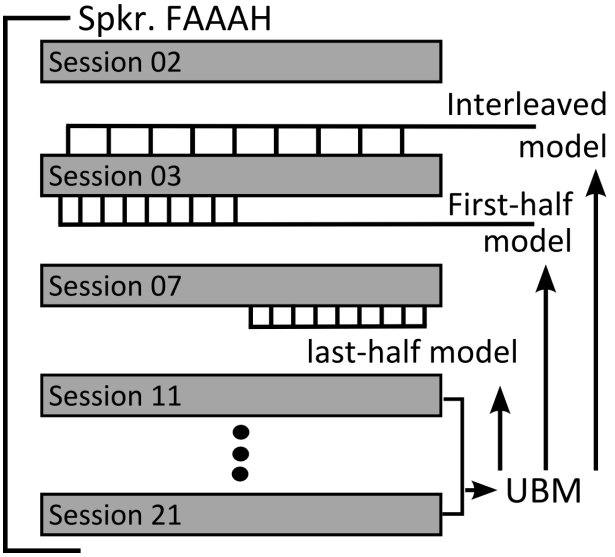

Figure 2: *The three training regimes of the second experiment, each model derived from a UBM formed only from sessions of the same speaker.*

5.1. Experiment 2: Description

In this experiment, 'claims' of session identity of 30-second chunks of speech are evaluated in a session verification paradigm, in which a model of the claimed session is compared to a model of several other sessions of the same speaker, a Universal Background Model or UBM. Three different choices of subsets of sessions to use as training data are investigated, in each case using 10 30-second chunks of each 10 minute session, with test data drawn from the 30-second chunks not employed in the modeling process. Figure 2 shows the 3 ways that 30-second chunks of sessions are grouped into training data sets for each session model: first half, last half, and interleaved. For each speaker, 7 sessions are modeled, with the remainder used to form a UBM. The acoustic features are 13 dimensional MFCCs with deltas and delta-deltas, and models are GMMs. Session models are Maximum A Posteriori (MAP) adapted from the respective speaker UBM.

The purpose of this experiment is realized by the varying of the subset of chunks used for training data. Comparing average performance resulting from the three subsets contrasts the relative amounts of intrasession and intersession variability observed in the corpus. True intersession variability that might support session verification may be due to microphone placement, or to long-term speech traits that have varied between other sessions, while the design of the MARP corpus precludes various non-speaker variations that often contribute to session-related performance problems observed in other corpora, such as change in microphone model or recording equipment, or changes in background noise. Conversely, intrasession variability not included in both the UBM and the session model may negatively impact the performance of session verification. Such an occurrence is more likely if session training data does not overlap the entire session, as in the case of the first half and last

Training regime	Mean EER	EER std. dev.
Train first half	21.10%	6.80%
Train second half	21.28%	6.69%
Interleaved train	20.04%	6.30%

Table 3: *Mean session verification EER across speakers for different train/test sets.*

Session	1	2	3	4	5	6	7
1	76.4	30.7	38.6	32.5	33.2	5.4	7.1
2	26.1	79.3	30.0	35.7	26.8	1.1	6.1
3	34.6	28.6	74.6	25.4	22.5	5.4	7.9
4	32.5	33.6	33.9	93.2	51.8	19.3	28.2
5	24.3	25.0	20.4	46.8	92.1	23.6	32.5
6	2.9	2.9	7.9	14.3	15.4	81.4	17.9
7	3.9	2.1	2.9	16.4	32.1	17.9	84.3

Table 4: *Confusion matrix for interleaved training regime. Each row shows the acceptance rates (in %) of the session in the row when the model in the column is presented as the claimed session. Thus, error rates are lower when the percentages on the diagonal are higher and those off the diagonal are lower.*

half training regimes here. Thus, a performance improvement observed in the interleaved training regime would indicate the occurrence of unmodeled intrasession variability.

Finally, to expand the available data used for analysis, two experiments are run in which the choice of which sessions to model is varied. In one run, the first 7 consecutive sessions available for the given speaker are modeled, with the remainder used to form the UBM. In another run, the last 7 consecutive sessions available for the given speaker are modeled. In figures that follow, session numbers 1 through 7 thus refer not to the first or second session recorded for that speaker, but to the first or second session in the sequence of modeled sessions.

5.2. Experiment 2: Results

Two perspectives on the results of the second experiment are shown in Tables 3, 4, and 5. Table 3 shows the average equal error rate (EER) for each training regime, averaged across speakers. Observed is a 1% absolute improvement in performance of the interleaved training regime over the first half and last half training regimes. Given the small size of the improvement, it is clear that identifying characteristics of the session are largely captured by training data from just the first or last half. However, the existence of the improvement shows that there is sufficient intrasession variability in the unmodeled half to corrupt performance.

Tables 4 and 5 show confusion matrices between modeled sessions, averaged across speakers, for the interleaved and first half train regimes. Due to space concerns, the confusion matrix for the last half training regime is omitted. It is nearly equivalent to the confusion matrix for the first half training regime, as is suggested by the nearly equivalent EER shown in Table 3. Each row of the confusion matrices indicates the acceptance rate associated with chunks from the listed session. Each column shows the acceptance rate associated with claims that a chunk is from the associated session. Thus, Table 4, position (1,2), shows that when chunks from the first modeled session are claimed as chunks from the second session, the acceptance rate of this claim is 30.7%.

Session	1	2	3	4	5	6	7
1	72.5	30.4	36.4	31.8	31.1	6.8	9.6
2	27.5	75.0	29.6	32.1	25.0	1.8	4.6
3	33.9	24.3	67.1	25.0	22.5	6.4	6.8
4	33.2	34.6	32.1	88.9	48.9	17.9	28.9
5	21.4	21.1	21.4	43.2	87.1	17.9	29.6
6	3.2	3.9	8.6	12.5	12.1	71.8	17.9
7	2.9	1.4	2.1	15.0	30.7	14.6	75.7

Table 5: *Confusion matrix for first half training regime. Each row shows the acceptance rates (in %) of the session in the row when the model in the column is presented as the claimed session.*

The confusion matrices show a combination of effects. An unidentified effect has separated the 6th and 7th modeled sessions from the other 5. On the other hand, within the first 5 sessions, the separation between sessions does not clearly increase with increased separation in time. Within these 5 sessions, this confirms the finding of [2] that aging is not a primary factor in intersession variability.

6. Discussion

Two experiments have been presented that shed new light on the relationship between intra-session and inter-session variability in speakers. The first experiment showed that intrasession variability dominates the relationships between 30-second chunks of speech, and that conversational tone forms one aspect of intrasession variability captured by this study's modeling technique. The second experiment showed that there are consistent differences between sessions that support session identification of 30-second chunks of speech, and that unmodeled intrasession variability contributes to performance degradation in session verification, but that it plays a minor role.

From these two experiments it may be concluded that medium-term variability in speech production patterns due to fatigue, mood, health, circadian rhythm, and other causes, contributes to performance problems traditionally associated with intersession variability, but that short-term variability in speech production dominates the variability observed in speech, obscuring the differences between sessions. One open question in speaker ID concerns the mechanisms that cause performance problems when testing on unseen sessions. This study has verified that there are long-term changes in speech patterns that identify a session, and that speech production changes are not necessarily related to progress through time. For future studies on intersession variability and speaker identification systems, this suggests that the relationships between sessions of a speaker need not be studied or modeled in chronological order, but rather that such studies may find more productive use in grouping sessions based on similarities in fatigue level, state in circadian rhythm, and other medium-term variability.

7. References

[1] A. D. Lawson, A. R. Stauffer, E. J. Cupples, S. J. Wenndt, W. P. Bray, and J. J. Grieco, "The multi-session audio research project (MARP) corpus: Goals, design and initial findings," in *INTERSPEECH 09*, 2009.

[2] A. D. Lawson, A. R. Stauffer, B. Y. Smolenski, B. B. Pokines, M. Leonard, and E. J. Cupples, "Long term examination of intrasession and inter-session speaker variability," in *INTERSPEECH 09*, 2009.

[3] S. Young *et al.*, *The HTK Book (for HTK Version 3.4)*. http://htk.eng.cam.ac.uk, 2006.

[4] S. Kullback and R. A. Leibler, "On information and sufficiency," *Annals of Mathematical Statistics*, vol. 22, pp. 79–86, 1951.