INTERSPEECH
2009 BRIGHTON

Long Term Examination of Intra-Session and Inter-Session Speaker Variability

A. D. Lawson[1]*, A. R. Stauffer*[1]*, B.Y. Smolenski*[1]*, B. B. Pokines*[2]*, M. Leonard*[3]**,** *E. J. Cupples*[1]

[1]RADC, Inc., Rome, NY USA
[2]Oasis Systems, Lexington, MA USA
[3]University of Texas, Dallas, TX USA

`Aaron.Lawson.ctr@rl.af.mil, stauffar@clarkson.edu, Brett.Smolenski.ctr@rl.af.mil, Benjamin.Pokines.ctr@rl.af.mil, mrl016000@utdallas.edu, Edward.Cupples.ctr@rl.af.mil`

Abstract

Session variability in speaker recognition is a well recognized phenomena, but poorly understood largely due to a dearth of robust longitudinal data. The current study uses a large, long-term speaker database to quantify both speaker variability changes within a conversation and the impact of speaker variability changes over the long term (3 years). Results demonstrate that 1) change in accuracy over the course of a conversation is statistically very robust and 2) that the aging effect over three years is statistically negligible. Finally we demonstrate that voice change during the course of a conversation is, in large part, comparable across sessions.

Index Terms: session variability, speaker recognition, speaker variability analysis, conversation analysis

1. Introduction

1.1. Session variability and speaker recognition

The three major goals of this project are: 1) how speaker variability changes within a conversation or session and what impact it has on automatic speaker recognition (SR), 2) how speaker variability changes over the long term (3 years) and its impact on recognition and 3) if there is a pattern of change in speaker variability with time that can be exploited to improve recognition performance. Intersession variability is a widely acknowledged problem for speaker recognition systems and the focus of much work on mitigation, e.g. [1]. There has been little investigation of whether inter-session variability is simply a function of changing conditions between sessions, or whether it reflects a continuum of changes that are already taking place as a conversation is on-going between participants. It is clear from the vast amount of research in the Conversation Analysis (CA) [2] community that voices change during the course of a conversation [3]. These changes reflect both incidental factors, such as emotion, engagement, empathy and fatigue, but are also due to procedural factors characteristic of initial contact, turn establishment, acclimation and termination and, as such, they tend to change in consistent patterns over the course of conversation. The effect of these changes on the voice and its relationship to the effectiveness of speaker recognition is a relatively unexplored issue, but an understanding of the impact or lack of impact can provide insight into improving speaker recognition and how speaker models fail.

While it is also widely assumed that as speakers age speaker recognition models will become less representative of the speaker, which *in extremis* is almost certainly true, the actual, shorter-term impact of aging is less than clear. This is probably due largely to a lack of good longitudinal data. Hébert [4], for example, assumes that aging is the cause of the significant loss in accuracy between two sessions separated by 3 months reported in Kato and Shimizu [5]. This loss in accuracy, however, is consistent with simple inter-session drop in accuracy as measured in [6] and other studies, and it cannot thus be unambiguously attributed to aging *per se*. For the purposes of this study inter-session variability is the manifest impact on a speaker's voice due purely to a different recording session, with no discernible change in channel, ambient noise, electromagnetic interference or other factors.

1.2. Goals of the project

The objective of this project is to investigate short term and long term speaker variability to improve intra-session and inter-session automatic speaker recognition performance. In examining these issues we will provide some clarification on the issues described in the literature and advance our understanding of these important factors that effect speaker recognition. It is essential to a project such as this one to lay out a rigorous statistical analysis of the dependence of speaker recognition performance and time over the period of a conversation (intra-session), between two sessions (inter-session) and across three years of sessions (aging) to demonstrate the significance of our findings.

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

The design of the MARP database allowed 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 address interest in the effects of time, aging, and intersession variability the MARP Corpus consists of multiple sessions of the same speakers recorded in 21 sessions over a three-year period of time. This study largely focused on 32 speakers in 672 sessions. Conditions were highly controlled, recordings were made in an anechoic chamber with consistent equipment and acoustic conditions throughout the three years.

3. The Speaker ID system

The Gaussian Mixture Model (GMM) and Universal Background Model (UBM) approach, developed by Reynolds [7], are used in this study. 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. For the implementation used in this paper only 5 mixtures are used for the calculation of the likelihood of a particular speaker's GMM model. The five mixtures are chosen from the most probable mixtures in the UBM. The accuracy of speaker recognition *per se* is not the basis of this study, rather we examine the impact of the experimental conditions on the log-likelihood scores output by the GMM speaker models. The notion being that, *ceteris paribus*, a lower log likelihood score indicates a lower match between a given model and a given segment of audio.

4. Approach to statistical analysis

The observed data is well modeled by the following two parameter power law:

$$y = ax^b \quad (1)$$

where a and b are the model parameters that are estimated from the data, x represents the time interval, and the response variable, y, represents the log-likelihood produced by the speaker identification system. Using this particular model makes intuitive sense, since one would expect the log-likelihood to decrease asymptotically over time as the conversation unfolded. By taking the natural logarithm of the independent and response variables:

$$\ln y = \ln a + b \ln x \quad (2)$$

one can approach the fitting of this nonlinear model using the familiar linear least-squares regression line [1].

The generalized correlation coefficient r measures how well a particular nonlinear model fits the observed data [2]. The value r^2, known as the *coefficient of determination*, corresponds to the ratio of the explained variation of the particular model to the total sample variation observed in the data:

$$r^2 = \frac{\sum (y_{est} - \bar{y})^2}{\sum (y - \bar{y})^2} \quad (4)$$

where y_{est} corresponds to the response values predicted by the model for observed value of x. The value of the coefficient of determination lies in the interval [0,1], with 0 corresponding to no correlation and 1 corresponding to total correlation, i.e. no error between the model and the data. In addition, the *standard error* $s_{y.x}^2$ of estimate of y on x gives a measure of the scatter about the regression curve and it is related to the coefficient of determination by:

$$s_{y.x}^2 = s_y^2 (1 - r^2) \quad (5)$$

with s_y^2 corresponding to the sample variance of the response variable.

Confidence limits for the population correlation coefficient ρ can be obtained by using the fact that the following statistic is approximately Gaussian distributed [3]:

$$Z = \frac{1}{2} \ln\left(\frac{1+r}{1-r}\right) \quad (6)$$

with mean μ_Z and standard deviation σ_z given by:

$$\mu_Z = \frac{1}{2} \ln\left(\frac{1+\rho}{1-\rho}\right) \text{ and } \sigma_Z = \frac{1}{\sqrt{n-3}} \quad (7)$$

Hence 95% confidence limits for ρ can be found by:

$$Z \pm 1.96\sigma_Z = \frac{1}{2} \ln\left(\frac{1+r}{1-r}\right) \pm 1.96\left(\frac{1}{\sqrt{n-3}}\right) \quad (8)$$

Different percent confidence limits can be obtained by using an alternative scaling of σ_Z.

5. Experiments

Speaker recognition experiments were run on the conversational part of the MARP Corpus to determine the extent of the impact of intra-session and inter-session variability. The major research questions were:

1) Change within a conversation: Does the accuracy of speaker ID change as one moves further in time from the model data within a conversation? If so, does it follow a significant trend, or does it change unpredictably?

2) Change over 3 years time: Does speaker ID accuracy degrade over time due to a speaker aging?

3) Matching position in a conversation: Is there a relationship between the position in a conversation one uses as training data and the accuracy of the model on various sections of other conversational sessions? In other words, is audio from the beginning of a conversation better at decoding the beginning of other conversations than it is at decoding other parts of the conversation?

5.1. Intra-session variability

This set of experiments looked at the effect of time within a conversation. Based on research by Goldberg [3] there was an expectation that changes in the voice measured in CA experimentation could impact speaker recognition and provide insight into session variability. Two conversational corpora were used for these tests 1) the MARP corpus, where 1015 conversations over 21 sessions were evaluated and 2) LDC's Call Friend corpus where a subset of 89 English conversations (178 speakers) were used to validate results. Data from each conversation were broken down into 30 second chunks, the first minute of each conversation was discarded, and the third chunk was used to train a GMM-UBM system. Each chunk thereafter was used as test data and the results where analyzed to determine whether distance within a single conversation impacted speaker recognition. To further validate results the tests were rerun with the temporally last chunk of data as the training chunk.

5.2. Impact of aging

Aging was evaluated on all 32 speakers who participated in the full range of 21 sessions in the MARP Corpus from June 2005 to March 2008. Session 2 was used as training data (session 1 was excluded from aging trials since it could be expected to differ from other sessions for other reasons). All other sessions were used as test data in 30 second chunks ("slice"). Analysis was performed to determine the impact of 33 months of aging on speaker recognition. Training on the last session (21) and testing on all others was done as a reverse validation.

5.3. Inter-session compatibility

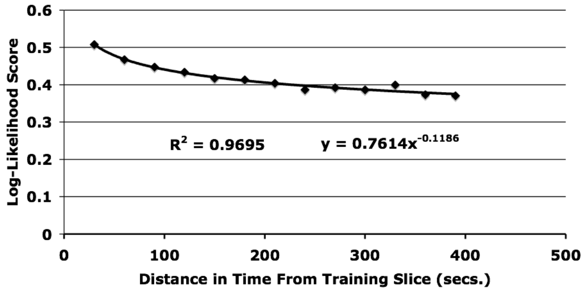
Further tests were performed to test question 3 above, with the goal of determining if a correlation exists between position in a conversation and effectiveness of training data. This was partially inspired by Goldberg's [3] research on regular and predictable shifts in talkers' voices during the course of a conversation. To verify whether there is an effect for position in a conversation models were generated from slice 3 in a given session and tested on 30 second slices from all other sessions.

6. Results

6.1. Intra-session variability results

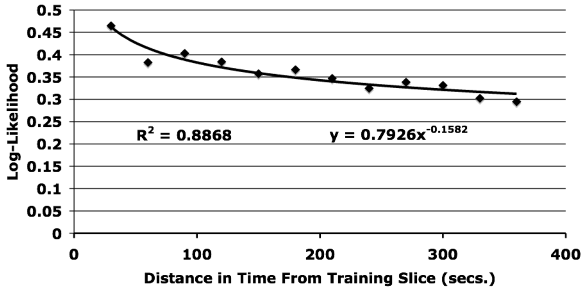
The most robust finding of this study was the strong correlation between time and speaker recognition log-likelihood scores for the target speaker. In the MARP corpus testing forward in time correlated along the power law curve with an r^2 value of .97, in table 1.

Table 1. *Change in log-likelihood scores as a function of distance in time from training (slice 3) MARP corpus*



Testing backwards in time yielded an r^2 value of .89, as can be seen in table 2.

Table 2. *Change in log-likelihood scores as a function of distance in time from training (backwards from slice 16) MARP corpus*

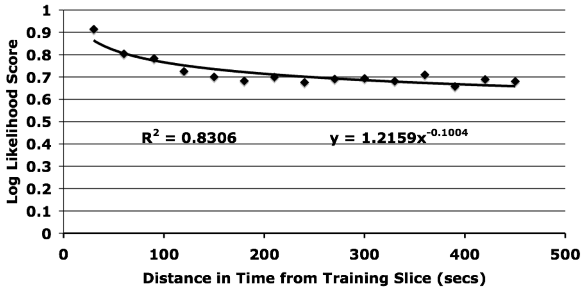


The validation set (Call Friend) provided an r^2 value of .83 (table 3).

6.2. Impact of aging

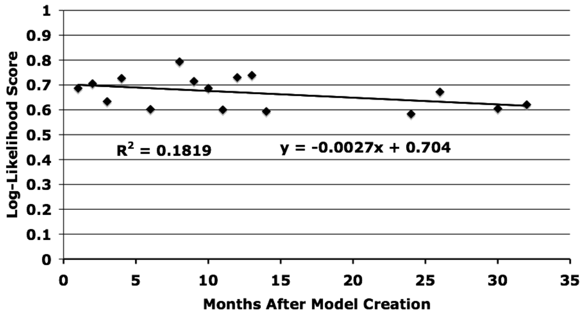
Aging was evaluated in a similar fashion, but statistical significance was measured by comparison to a linear correlation with time due to the fact that there is no expected "tapering off" of changes to the voice due to aging, while one would expect the changes in speaking during a conversation to "bottom out" at some point.

Table 3. *Change in log-likelihood scores as a function of distance in time from training (Slice 3) Call Friend corpus*



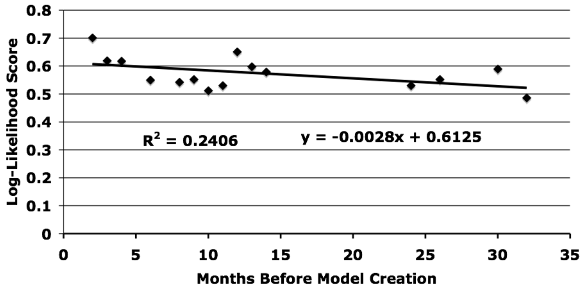
While speaker recognition is strongly impacted by testing on a different session, this study found only a very slight correlation with time.

Table 4. *Longitudinal impact of time (1-32 months from session 2) on log likelihood scores-MARP Corpus*



As table 4 above shows the r^2 value of speaker recognition log-likelihood scores with time is only .18 when tested going forward in time. Testing backward in time yields a slightly higher correlation, with a r^2 value of .24, in table 5.

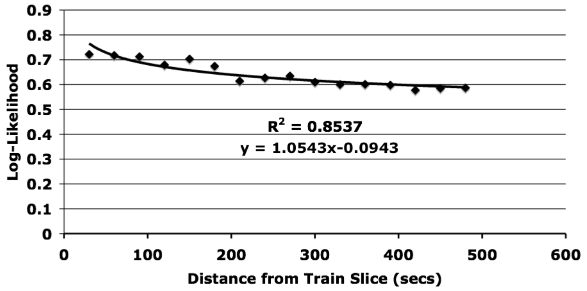
Table 5. *Longitudinal impact of time (1-32 months before session 21) on log-likelihood scores -MARP Corpus*



6.3. Inter-session Compatibility results

Using training slice three from one session to decode other sessions manifests a pattern similar to that discussed in section 6.1, showing that position in a conversation correlates with accuracy of a model.

Table 6. *Log-likelihood scores as a function of distance in time from training slice across sessions (train on slice 3) MARP corpus*

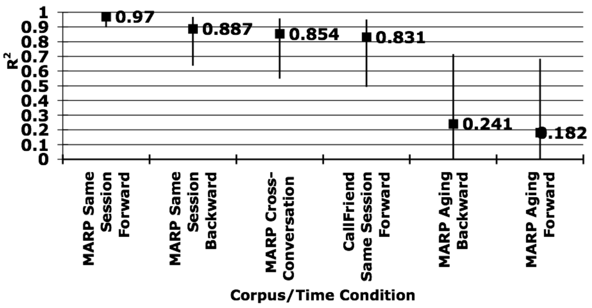


As one can see in table 6, effectiveness of model correlates with position in a conversation with an r^2 value of .85

7. Discussion/Conclusions

The most important conclusions of this study are 1) that the inter-session degradation observed in [6] and elsewhere is already occurring within a conversation. This result has the highest statistical significance of all findings in this study and the tightest confidence range. This was verified by testing backward in time and evaluating another corpus. 2) Position in a conversation correlates with efficacy of a model, i.e. models trained with data from the beginning of a conversation perform better on data from the beginning of another conversation than data from the end. The most surprising finding is 3) the lack of significant progressive degradation in speaker recognition over the course of the three years of this study. In fact, while the impact on speaker recognition accuracy between any two sessions is considerable, the long-term trend is statistically quite small, and has a very large confidence variance. Indeed, when one compares the r^2 and confidence range of research questions 1 and 2 with question 3 in table 7, it is clear that the "aging" effect across the 21 sessions of this study is minor.

Table 7. *Correlation of Determination (R^2) and confidence ranges for speaker recognition conditions tested*



This result clarifies the findings in Hébert [4] and Kato and Shimizu [5] that there is indeed a detrimental impact on recognition accuracy across sessions, but that it is clearly not primarily a function of aging or of the voice changing within this timeframe. Further we see this process of model degradation occurring within the same conversation. This lends support to the finding of the CA field that important characteristics of a speaker's voice, more related to "register" fluctuations than to permanent changes in a speaker's voice, are crucial factors in inter- and intra-session variability.

7.1. Implications of this study for speaker recognition

Foremost, awareness of the impact of the kinds of session variability examined in this study is a very important step towards understanding the factors that affect speaker recognition, factors that are clearly separable from channel, noise and physical environment. A preliminary follow-on study has found significant correlations between intra-session degradation and several modal voice characteristics, including average amplitude, average voiced segment energy, and formant frequencies of F2 and F3. While these voice factors are probably not directly responsible for the impact on speaker recognition they are a first step towards understanding what is happening to the voice over the course of a conversation and between sessions. In addition, he findings of this project have already formed the basis to an approach to mitigating session variability for speaker identification in [11]. Future research will logically focus on further understanding the modal and non-modal aspects of the voice that correlate with the observed phenomena, and proceed to additional mitigation strategies and improvements in the robustness of speaker recognition based on these findings.

8. References

[1] Stolcke, S. S. Kajarekar, L. Ferrer, and E. Shriberg, "Speaker recognition with session variability normalization based on MLLR adaptation transforms", IEEE Trans. Audio, Speech, and Lang. Process., vol. 15, pp. 1987–1998, Sep. 2007.

[2] Schegloff, Emanuel, *Sequence Organization in Interaction: A Primer in Conversation Analysis, Volume 1*, Cambridge: Cambridge University Press, 2007.

[3] Goldberg, J. A. "The Amplitude shift mechanism in conversational closing sequences", in *Conversation Analysis: Studies From The First Generation*, Gene H. Lerner (Ed.), John Benjamins Publishing Company, 2004.

[4] Hébert, M. "Text-Dependent Speaker Recognition," *Springer Handbook of Speech Processing*, Springer-Verlag: Berlin, 2008.

[5] Tsuneo Kato and Tohru Shimizu, "Improved Speaker. Verification Over the Cellular Phone Network Using Phoneme-Balanced and Digit-Sequence Preserving Connected Digit Patterns" Proc. IEEE ICASSP 2003.

[6] 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

[7] 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.

[8] Cuthbert D. and Wood F.S., *Fitting Equations to Data: Computer Analysis of Multifactor Data*, 2nd ed., Wiley-Interscience, 1999.

[9] Devore J.L., *Probability and Statistics for Engineering and the Sciences*, 7th ed., Duxbury Press, 2007.

[10] Draper N.R. and Smith H., *Applied Regression Analysis*, 3rd ed., Wiley-Interscience, 1998.

[11] 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.