9. Detection of Demographics and Identity in Spontaneous Speech and Writing

A. Lawson, L. Ferrer, W. Wang and J. Murray

Abstract. This chapter focuses on the automatic identification of demographic traits and identity in both speech and writing. We address language use in the virtual world of on-line games and text entry on mobile devices in the form of chat, email and nicknames and demonstrate text factors that correlate with demographics, such as age, gender, personality and interaction style. Also presented here is work on speaker identification in spontaneous language use, where we describe the state of the art in verification, feature extraction, modeling and calibration across multiple environmental conditions. Finally, we bring speech and writing together to explore approaches to user authentication that span language in general. We discuss how speech-specific factors such as intonation, and writing-specific features such as spelling, punctuation and typing correction correlate and predict one another as a function of users' sociolinguistic characteristics.

9.1 Introduction

This chapter investigates several facets of how identity and demographic categories are manifested in spoken and written language use, along with approaches to capturing this information for real world analysis, authentication, and talker and writer identification. The first section details work done by the VERUS (Virtual Environment Real User Study) team under the AFRL (Air Force Research Laboratory) VERUS program, which was tasked with identifying features in virtual world activity that contribute to predicting the real world demographics of the participants involved. In this chapter we specifically focus on virtual world language use, which generally came from two sources: on-line chat and avatar nick-names. This work was crucial to providing features to determine gender, age

This material is based on work supported by the Defense Advanced Research Projects Agency (DARPA) under Contract D10PC20024. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the view of DARPA or its contracting agent, the U.S. Department of the Interior, National Business Center, Acquisition & Property Management Division, Southwest Branch.
This material is based on work supported by United States Air Force and DARPA under Contract No. FA8750-13-C-0280. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the United States Air Force and DARPA.
Approved for Public Release, Distribution Unlimited

group, ethnicity, education level and nativeness of the real world participant based solely on activity in the virtual world.

The next section switches focus to spoken language, and the recent progress that has been made in the domain of speaker identification from their voice. We focus on the major problems inherent in speaker identification, both differences inherent to the talker (language, phonetic content, speaker state) and external factors (channel of collection and transmission, noise, reverberation). We examine the recent findings in terms of features (acoustic and prosodic), as well as modeling techniques that have provided breakthroughs in recent evaluations, such as low-dimensional iVector representations of an utterance and probabilistic linear discriminant analysis (PLDA) for score generation. Further, we discuss the important area of calibration, in particular the issue of maintaining a coherent representation of the likelihood of a speaker given a specific utterance across a range of varying conditions.

The final section presents on-going work that combines research from both written and spoken authentication and characterization approaches under the DARPA (Defense Advanced Research Projects Agency) Active Authentication program. The goal of this work is to provide continuous authentication of users on their mobile devices using spoken and written inputs on the device, such that if an unauthorized user accesses the device their behavior will quickly reveal them to be an unauthorized user. This continuous authentication will make use of the shared space of language, which covers speech and writing, and the sociolinguistic relationships that emerge from the intersection of language use and personality, background, gender, age, ethnicity, interaction style, etc. Our ultimate goal is to develop a framework for predictive models of users that is robust to incomplete enrollment samples, making use of natural feature correlations across speech and writing.

9.2 Demographics in Virtual World Environments

9.2.1 Background

This section focuses on research into the relationship between virtual world or on-line linguistic behavior and real world demographic characteristics, with the goal of automatically predicting major real-world (RW) demographic attributes using only virtual world (VW) behavior. Much of this research came out of the VERUS study (Dieterle and Murray 2011). The RW attributes studied include age group, gender, ethnicity, income level, education level, leadership role and urban/rural background, among others. Volunteer participants provided their RW demographic information and allowed their on-line behaviors to be recorded. Over one thousand participants generated data during on-line activities, including text chat and names chosen for on-line personae (aka their "avatars"). Hypotheses were gathered from the theoretical sociolinguistics literature, phonology and sound symbolism, semantics and discourse analysis and from empirical observations of the data collected to generate features. These features were combined in a global model using statistical classifiers that enabled high-accuracy prediction of users RW attributes.

Table 1: Data Distribution of the VERUS Corpus

Game	*Turns*	*Talkers*	*Tokens*
Guardian Academy	914	57	2688
Sherwood	13,149	271	57,843
SecondLife	79	4	392
WoW	2337	117	56,036
Total	*11214*	*445*	*89,521*

Participants ranged from minors in their early teens to retirees in their 70's from Canada, the United Kingdom and the United States. Data from four virtual worlds was collected from existing on-line communities, namely SecondLife and World of Warcraft, and two VWs that were specifically developed for this study, Sherwood and

Guardian Academy (see Murray et al. 2012).

9.2.2 Features

The focus of feature development was both to understand what factors and behaviors manifest in the text were associated with specific demographic categories and to identify textual elements that could be automatically extracted for use in effective machine learning. A substantial amount of feature research involved understanding the motivation behind a phenomenon in the text (e.g. use of ellipsis) and its association with a demographic category (older users). The primary sources for features were thus identified using both a top-down and bottom-up approach. The top-down features were motivated from findings in the socio-linguistic literature –claims about how males and females used language differently, or how adult language use would differ from teenaged language use, etc. Bottom-up features arose from an examination of the data itself. This is especially important since virtual world and on-line discourse represent emerging modes of communication and there is a reasonable expectation that theories of traditional spoken and written discourse may be inadequate in this context.

Sources for top-down features mainly came from studies of gender and discourse, beginning with Robin Lakoff's work in 1975, and including studies by Deborah Tannen (1984, 1994), Shuttleworth and Keith (S&K) (2000), and O'Barr and Atkins (1980). Susan Herring (1994, 2006) has specifically focused on the interaction between language and gender in on-line communities yet many of her findings corroborate the results of the earliest studies by Lakoff. In general, the literature points out that linguistic features associated with females tend to be attenuative, indirect and cooperative, while male linguistic behavior tends to be more adversarial, direct and independent. Table 2 summaries the major traits identified in the literature that were investigated in this study.

Table 2: Gender traits from the socio-linguistic literature

Trait	*Gender*	*Source*
Hedging, hesitation, uncertainty	F	Lakoff, Herring
Polite forms	F	Lakoff
Challenging or confrontational forms	M	Herring
Question forms and intonation	F	Lakoff, S&K, Herring
Frequent or gratuitous apologies	F	Lakoff, Herring
Modal verbs	F	Lakoff
Insults, cursing or put downs	M	Lakoff, Herring
Contentious assertions	M	Herring
Supportive and empathetic statements	F	S&K
Sarcasm, self promotion	M	Herring
Agreeing and thanking	F	Herring
Commands	M	S&K

In addition, new phenomena that were observed to correlate highly with demographic classes were also added to our set of features (see table 3). These include typographic variations, which are associated with age differences, as well as more subtle distinctions between slurs and direct and indirect apologies.

Additional features came from consultation with Subject Matter Experts (SME) on virtual worlds, freely available lexical class databases, and features developed from the structure of virtual world environments. We divide these features into two sets: lexical features and structural features.

Lexical features include unigram probabilities of words in the list of Internet slang and emoticons, unigram probabilities of other non-standard words/short-hands (e.g., *ur (you are), thn (then), im (I'm), qust (quest*)), features related to disfluencies, features related to person addressing, and features related to occurrence of foreign characters.

Table 3: Socio-linguistic features from data observations

Trait	*Demographic*	*Example*
Use of slurs	Male	"You jerk!"
Direct apologies	Female	"I'm sorry"
Indirect apologies	Male	"Ooops", "my bad"

Standard Emoticons	Female	:) : (
Use of all caps	Youth	"STOP BEING DUMB"
Frequent use of ellipsis	Adult	"if you bring up your questlog..."
Commas, apostrophes	Adult	"we're done. let's turn in"
Lowercase 'i' for 'I' and 'u' for 'you'	Youth	"u losted to 4 pokemno"
Single word texts	Youth	"come", "yo"

We also studied features representing sentence complexity and structure, including vocabulary size of a participant, maximum length of 10% most frequent words, features related to discourse markers, and features related to grammaticality (estimated based on the normalized likelihood of the sentence from a state-of-the-art statistical English parser adapted to game text chat).

We further used features from databases of lexical categories, including the "Dictionary of Affect in Language" (DAL) (Whissel 2009) and the Linguistic Inquiry Word Count (LIWC) (Pennebaker et al. 2007) The DAL is an instrument designed to measure the emotional meaning of words and texts. It does this by comparing individual words to a list of 8,742 words that have been rated by people for their activation, evaluation, and imagery. Each word in the lexicon also receives a score according to "pleasantness", "activity", and "imagery". Then we computed the average of these scores for the sentences contributed by a participant and average counts of words belonging to each of these three categories, and used them as DAL features. The goal of LIWC was to identify a group of words that tapped basic emotional and cognitive dimensions often studied in the social sciences, health and psychology and use them as features reflecting disposition, personality, etc.

We also developed a set of "structural" features, inspired by conversation analysis and based on observations from game subject matter experts on player behavior, to quantitatively represent the different participation patterns of participants in a text chat conversation. Structural features for a participant include the percentage of sentences and 'turns' (i.e. exchanges between speakers in a conversation) from that participant out of all sentences and turns in the session respectively, and the average number of words per sentence and per turn of that participant.

We also studied features related to "structure" of game chat. For example, "silence" (no chat) durations from a participant in a game session may be synchronized with other gaming activities (i.e., the participant was possibly busy with other in-game activities hence couldn't contribute much text chat). Other structures that were useful were the use of positive and negative extreme case formulations –for example "that was the worst game ever" and expressions of self-affirmation (e.g., *I'm the best*). We further extracted features based on social network analysis, by capturing who is talking to who and how frequently they do it in the text chat.

An additional set of features was also developed based on the names users chose for their avatar and VW character. Since the name chosen by a particular player generally reflected information about the player in terms of gender, age, personal interests, ethnicity, etc. many effective features were extracted from components of the avatar names. This included sound symbolism-based features (Ohala, et al. 1994) related to gender (e.g. female names having high, front vowels, sibilants and ending in 'a'), typographic features related to age (use of capitalization, numbers and special characters), and features based on real world cultural references in the name (e.g. youth names referencing elements from Harry Potter books). For a more detailed overview of the avatar naming phenomena see Lawson and Taylor (2012) and Lawson and Murray (2013).

9.2.3 Machine Learning and Findings

We further explored machine learning approaches to identify predictive features from a variety of features, for detecting real world (RW) target variables based on virtual world (VW) data. We applied these techniques to predicting the following RW target variables:

gender, age group, community, ethnicity, English nativeness, RW and VW leadership and followership.

Two classification approaches were found to perform well on the features extracted from text chat from players: AdaBoost and linear kernel Support Vector Machines (SVMs), using RW target variable labels for supervised training. For feature normalization, we compared the effect of mean/variance normalization and rank normalization on RW target variable prediction accuracy. For feature selection, we compared forward-backward feature selection, the SVM-RFE (Support Vector Machine-Recursive Feature Elimination) algorithm, and logistic regression for fusion parameters.

We evaluated precision, recall, and overall classification accuracy for target variables, from 10-fold cross-validation for model training/testing and feature selection, by polling text chat from all four VW together. We built multi-class classifiers for predicting RW target variables. For example, for predicting age group, the overall classification accuracy is 82.54%; for predicting community, the overall classification accuracy is 83.51%. The most predictive features for age and gender are presented in tables 4 and 5, below.

Table 4: The most predictive features for age group (Adult >24, YoungAdult (18-24), Youth (<18)

Rules for Age Group
• Adult has larger average number of words per turn than Youth
• Youth use all uppercase (shouting) much more than non-youth
• Adult has larger average modal words per sentence than Youth
• Youth has larger average number of name addressing per sentence than YoungAdult
• Youth has larger average number of disfluencies per sentence than Adult
• YoungAdult uses more Internet slang per sentence than Youth
• Youth tend to use single word utterances at a rate almost double than adults and to use shorter phrases in general
• YoungAdult uses more extreme case formulations than Youth
• The use of proper contractions (e.g. “I’m” vs “im”) increases with users’ age

• The use of both periods '.' and commas ',' increases with users' age
• Adults' use of personal pronouns is double that of Youth
• Adults are twice as likely to use traditional emoticons as Youth and 3-4 times more likely to use ellipsis

Gender conclusion both tend to corroborate the findings of the socio-linguistic community while adding new findings to our understanding of how gender traits manifest in language use.

Table 5 Most predictive features for gender

Rules for Gender
• Females tend to use hedging forms more than men, including modal verbs, expressions of uncertainty and questions
• Males tend to use more offensive language and slurs than females, though females use more attenuated swears (e.g. 'darn', 'crud')
• Females tend to apologize more than males, though males use more indirect apologies ('Ooops')
• Females are more likely to agree and to express empathy than males
• Females are more likely to use traditional emotions than males, but males are more likely to use lewd emotions
• Both females and males choose avatar names or nicknames that tend to conform to the findings of the sound symbolism literature (Jespersen, Ohala, Gordon and Keith)
• Female avatar names tend to end in 'a', have sibilant consonants ('sh') or front vowels 'i, y, e'
• Male avatar names tend to end in back vowels ('u', 'o') or consonants, especially back or alveolar stops

9.2.4 Combined results

Final results were obtained by combining the language-based features with features from VW economic activity, movement, dress and game play activity (e.g. dueling). Combination was done using the WEKA toolkit (Hall et al. 2009) and feature selection techniques based on work by Flach and Lachiche (2001). The main goal of the program was arriving at human-understandable and human-usable rules –thus precision for each category and each feature was of paramount importance, with recall being of lesser importance. Results for the main categories are presented in table 6, below.

Table 6: Final combined results for the major program demographic targets

Category	*Precision*	*Recall*
Gender	98%	52%
Approximate age group	88%	13%
Ethnicity	83%	37%
English as Native Language	77%	22%
Education	79%	52%
Socioeconomic Status	83%	67%
Income Level	85%	35%

Precision was prioritized over recall in this program, since the overall goal was to develop human-understandable collections of rules that could effectively pick out demographic traits with high accuracy. Thus, even if a rule only applied to a minor sub-set of the total population, so long as it was precise, it was of high utility.

9.3 Detecting Identity in Large Collections of Spontaneous Speech

Automatic speaker recognition is the task of recognizing the person speaking in an audio recording. It can be classified into two main tasks: identification and verification. Speaker identification aims at identifying the speaker present in the recording among a set of known speakers. Speaker verification, sometimes also called speaker detection, on the other hand, aims at deciding whether the audio recording corresponds or not to a certain speaker of interest.

The task of speaker identification can be solved through a series of speaker verification queries against all target speakers, as explained for the related tasks of language detection and identification by Brummer (2006). Furthermore, given its binary nature, verification is easier to define and evaluate than identification. For these reasons most research in the area has been done for the speaker verification task.

Speaker verification is used in security applications to verify whether a speaker is who he claims he is. It is also used in applications that search for specific speakers within a large database of speech. In the last few years, speaker verification performance on clean telephone data has reached extremely good performance levels, with error rates below 1% for recordings of around 2 minutes of duration making the technology adequate for use under these conditions.

On the other hand, performance on harder recording conditions involving noise, channel distortion, reverberation, and other non-ideal conditions is severely affected and can reach unusable levels. Nevertheless, much progress has been made in these areas in recent years. This section covers some of the techniques that have lead to major improvements under these challenging scenarios.

9.3.1 Overview

The core speaker verification task is defined as determining whether a specified target speaker is speaking during a given segment of speech. More explicitly, one or more samples of speech data from a speaker (referred to as the "target" speaker) are provided to the speaker recognition system. These samples are the "training" or "enrollment" data. The system uses these data to create a "model" of the target speaker's speech. Then a sample of speech data is provided to the speaker recognition system. This sample is referred to as the "test" segment. Performance is judged according to how accurately the test segment is classified as containing (or not containing) speech from the target speaker.

Metrics that reflect accuracy are related to a typical hypothesis test (i.e., based on false positives (referred to as false alarms) and false negatives (misses)). In this work, we report equal error rates

(EER), where false alarm and miss rates are equal, or the false alarm rate at a particular miss rate. The performance of a system over the range of possible operation points is generally represented in a Decision-Error Tradeoff (DET) curve, which plots the relationship between false alarms and misses over all points.

9.3.1.1. Challenges

As for any detection task, the main challenge of speaker recognition is extracting features that will represent a speaker independent of variations that can occur in the observations. Minimizing the intra-class variability while maximizing the inter-class variability is our goal.

Speech is a complex signal, and many possible variations of that signal exist for the same individual. During the previous few years, the community has tackled the problem of extrinsic variability and how to factor out extrinsic variability from the speaker model (sometimes referred to as channel compensation in articles). This kind of variability is detrimental to high accuracy speaker recognition. Indeed, recorded speech varies as a function of many factors that are not a function of the speaker's identity, including: acoustic environment (e.g., background noise), channel (e.g., microphone, handset, recording equipment), high signal-to-noise ratio (SNR), audio degradation through compression, speaker's physical condition (emotion, intoxication, illness), what is said (text-independent versus text-dependent), and speaking context (level of formality, planning, language).

9.3.1.2 Approaches for Mitigation of Undesired Variability

Mitigation of the undesired variability can be performed at different levels in the system from the feature extraction to the last stage of system fusion. This section summarizes several different approaches, focusing on recent methods implemented in state of the art systems.

1) Feature Diversity

A successful approach to speaker verification is to combine different knowledge sources by separately modeling them and by fusing them at the score level to produce the final score that is later thresholded to obtain a decision. Combinations of systems are most successful when the individual systems being combined are significantly different from each other.

Prosody—the intonation, rhythm, and stress patterns in speech—is not directly reflected in the spectral features. As a consequence, these features show great effect in combination with traditional features (Kockmann, 2011). The state-of-the- art approach to extracting prosodic features is to compute the pitch and energy contour in the signal using Legendre polynomial coefficients.

Standard spectral-based features include perceptual linear prediction (PLP) features and mel-frequency cepstrum coefficients (MFCC). In addition, many spectral-based features were developed specifically for noise-robustness under the DARPA RATS (Robust Automatic Transcription of Speech) program. Medium duration modulation cepstrum (MDMC) features (Mitra, 2012) extract modulation cepstrum-based information by estimating the amplitude of the modulation. Power-normalized cepstral coefficient (PNCC) (Kim, 2012) features use a power law to design the filter bank as well as a power-based normalization instead of a logarithmic one. Mean Hilbert envelope coefficient (MHEC) features (Sadjadi, 2011) use a gammatone filter bank instead of the Mel filter bank, and the filter bank energy is computed from the temporal envelope of the squared magnitude of the analytical signal obtained using the Hilbert transform. Subband autocorrelation classification (SACC) (Lee, 2012) provides a pitch estimate from an estimator that is trained using a multilayer perceptron, allowing for a robust prosodic system implementation, which we call PROSACC (Prosodic Subband AutoCorrelation Classification) in this article.

2) Advanced Modeling

Recently, the speaker-verification community has enjoyed a significant increase in accuracy from the successful application of the factor analysis framework. In this framework, the i-vector extractor paradigm (Dehak, 2010) along with a Bayesian backend is now the state-of-the-art in speaker verification systems. An i-vector extractor is generally defined as a transformation where one speech utterance with variable duration is projected into a single low-dimensional vector, typically of a few hundred components.

The low rank of the i-vector itself opened up new possibilities for the application of advanced machine-learning paradigms that would have been otherwise too costly with the very high dimensionality used by most earlier systems. Probabilistic linear discriminant analysis (PLDA) (Prince, 2007, Kenny, 2010) has proved to be one of the most powerful techniques for producing a verification score. In this model, each i-vector is separated into a speaker and a channel part, analogous to the formulation in the Joint Factor Analysis framework (Kenny, 2008).

A simple and quite effective approach for robustness against undesired variability is to include data with the corresponding variability during training of the PLDA model (Lei, 2012), so that the model can learn the appropriate intra-speaker variability under the conditions of interest. On the other hand, the components of the i-vector extractor seem much less sensitive to exposure to a new type of variability and can be kept untouched without much, if any, effect in performance.

3) Metadata Extraction

Metadata information about the audio recording can be used to affect the parameters of the models, allowing adaption to the specific conditions of the recording. Rather than relying on either annotated data, or developing specific systems for each type of variability, a universal audio characterization system can be used to extract metadata information based on the i-vector (Ferrer, 2010). This enables the system to detect if an audio recording contains certain kinds of noise, channels, or the speaker's gender or language.

4) System Fusion and Calibration

Fusion of systems is usually performed either at the score level or at the i-vector level. At the score level, system fusion is generally performed using logistic regression with a cross entropy objective (Brummer, 2007), the standard fusion approach in speaker recognition. This approach offers the benefit of producing calibrated scores, treatable as log-likelihood ratios, which are ideal for forensic comparisons and decisions.

As mentioned in (Ferrer, 2010), the metadata extracted from the universal audio characterization system can be used during fusion to adapt the output score to the signal's conditions. A modified version of the logistic regression fusion algorithm is used so that log-likelihood ratios are still produced but are biased depending on the metadata between the enrollment and test utterances.

9.3.2 Robustness to Undesired Variability

In this section, we highlight the impact of the approaches described above for different types of degraded audio conditions and other extrinsic variations.

9.3.2.1 Channel, Noise, Reverb, Vocal Effort and Language Variation

To evaluate speaker recognition accuracy on multiple types of variability, SRI created the PRISM (Promoting Robustness for Speaker Modeling) dataset (Ferrer, 2011), building on data previously collected by the Linguistic Data Consortium and creating trials from waveforms degraded by adding noise or reverberation. The PRISM data set is available online at https://code.google.com/p/prism-set/.

In Figure 1, we show the benefit of different mitigation approaches by showing the increase in speaker recognition accuracy for every step of the pipeline. Note that results for different conditions are not comparable since they involve different speakers and other factors that affect the absolute performance. Comparisons should be made within condition and across systems.

The conditions defined in the PRISM set and represented in the horizontal axis of the figure are:

- *telphn:* telephone calls over telephone channels
- *intmic:* microphone recordings in an interview setting
- *telall:* telephone calls over telephone channels and other microphones
- *voc:* vocal effort: low and high
- *lang:* Trials made of languages other than English

- *noise:* Clean signals degraded with real noise samples at different SNR levels ranging from 20 dB to 6 dB.
- *reverb:* Clean signals degraded with artificial reverb at reverb times (RT) of 0.3, 0.5, and 0.7 seconds

The baseline system is a standard i-vector/PLDA recognition pipeline on MFCC features, without the mitigation mechanism for the variations of interest.

The robust system uses an enhanced PLDA model designed to be robust to the variations of interest by adding data with these types of variation during training. This system also includes other techniques that add robustness to the system, namely i-vector length normalization (Garcia-Romero, 2011), an LDA step for dimensionality reduction, and i-vector adaptation, where the mean i-vector over each condition is subtracted from the corresponding i-vectors before PLDA modeling. Improvements are highly significant, reducing error by a factor of 10 times on the noise condition while also improving results for "cleaner" conditions like telephone calls.

The robust + prosody system is a fusion of the robust MFCC system and a robust prosodic system. We see that an additional improvement can be observed in most conditions. Finally, we enable metadata extraction and handling in the robust + prosody + metadata system to obtain additional improvements for most conditions except language and vocal effort. These conditions were not represented as classes for the metadata extractor due to lack of training data for them and, hence, could not be appropriately predicted.

The figure shows how the different approaches for mitigation reduce the effect of the undesired variations, in some cases reducing the false alarm rates at 10% miss rate by an order of magnitude.

9.3.2.2 Highly Degraded Channels

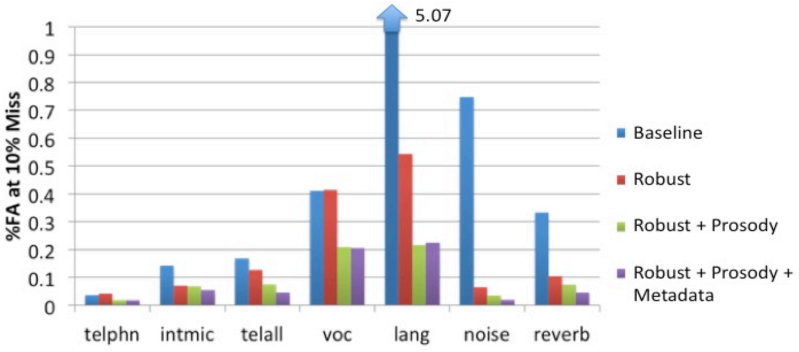


Figure 1 SRI's speaker verification results on the PRISM set.

The DARPA RATS program aims at developing robust processing methods for speech acquired from highly degraded transmission channels. The audio recordings (Walker and Strassel 2012) used in the RATS program are severely degraded with additive noise, channel-convolved noise, bandwidth limitations, and frequency shifting. Telephone conversations are retransmitted over eight different military transmitter/receiver combinations. All the data was retransmitted across all the channels and re-recorded, resulting in more than 100,000 files. The core languages from which speakers are selected are Levantine Arabic, Farsi, Dari, Pashto, and Urdu. In the speaker-verification task each speaker model was trained using six different sessions from different channels. A trial was designed using one speaker model and one test session. The duration of each training and testing session was 3, 10, 30 or 120 seconds depending on the condition.

SRI's system was composed of five different features: PLP, MDMC, MHEC, PNCC, PROSACC. For the i-vector framework used by all feature streams, we used universal background models (UBMs) with 2048 diagonal covariance Gaussian components trained in a gender-independent fashion. The PROSACC systems used 1024-component UBMs. The i-vector dimensions of 400 were further reduced to 200 dimensions by LDA (in the case of PROSACC, 200D i-vectors were reduced to 100D), followed by length normalization and PLDA.

The systems are fused at the i-vector level by concatenating each i-vector from each stream into a single vector before employing the PLDA backend. The i-vector dimensions are first reduced using

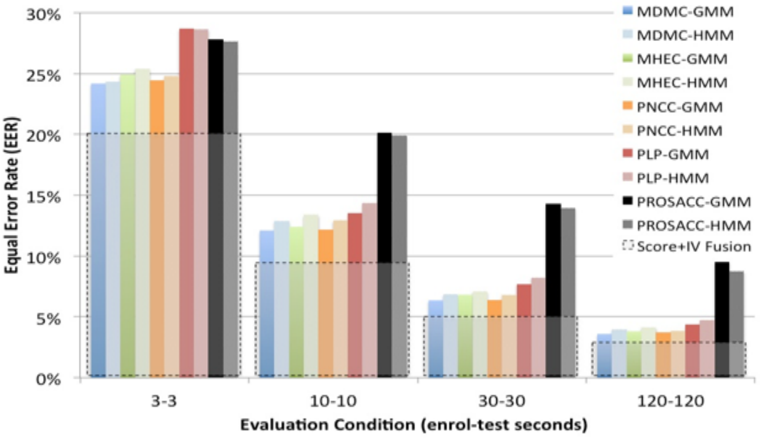


Figure 2 SRI speaker recognition system results on the DARPA RATS development set for different combinations of train and test durations.

LDA, and only after concatenation does a second dimensionality reduction shrink the total dimension to 200. Fusion of systems at the score level was performed using logistic regression.

Results from four core conditions are provided in Figure 2 above, showing the relative performance of the five acoustic features with both HMM (Hidden Markov Model) and GMM (Gaussian Mixture Model) SAD (Speech Activity Detection), as well as the gain from the final score plus i-vector fusion system (in dashed lines). For more details on the results presented in this figure, see (McLaren 2013). For all durations, the MDMC and PNCC features with GMM SAD had the least errors. The fusion system was always significantly better than any single system, benefiting in particular from the PNCC features and substantially from the inclusion of PROSACC, despite the system's low accuracy on its own.

9.4 Work in Progress: Identification of identity factors from both speech and text in the Active Authentication program

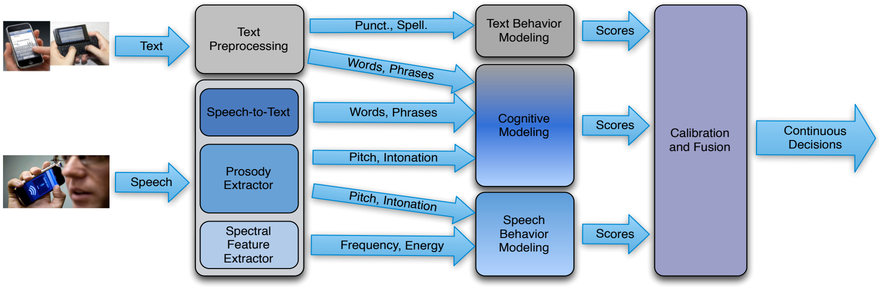
This section describes an on-going project, LinguaKey, to combine both speech and text data for continuous authentication of mobile device users through their language usage. The goal of this work system is to provide continuous authentication of users actively performing routine tasks on a mobile device. These include telephone conversation, spoken device interaction (e.g., "SIRI"), text chat, instant messaging and email. This research builds a profile of a user's distinctive linguistic usage and generates models to provide an ongoing means of ensuring that only authorized individuals have access to their mobile device.

Our approach is to identify discriminative features based on ***behavioral*** and ***cognitive*** dimensions of individual linguistic variability. These factors represent deeply ingrained parts of a user's way of thinking and behaving, of which the user may not even be conscious. The notion is that since these factors are difficult to control they will be resistant to spoofing. Behavioral dimensions include features from speech and text production, i.e. the idiosyncrasies of how users actually produce spoken or written language. In speech this includes factors such as *intonation*, *speech rate*, and *speech energy/frequency*. In text, behavioral features will target *spelling errors*, *punctuation style*, use of *abbreviations*, *sentence* and *word length,* among other factors. Cognitive features are focus on sociolinguistic factors derived from linguistic activity that provides clues about the individual user's background, personality and approach to interpersonal relations. An individual's language use is highly colored by sociolinguistic background, with influences from gender, age, ethnicity, native language, region, level of education, social class, dialect and others. We further leverage the rich information in word sense information and linguistic context to provide parameters for identifying users based on personality traits and their typical approaches to interacting with others.

9.4.1 Approach

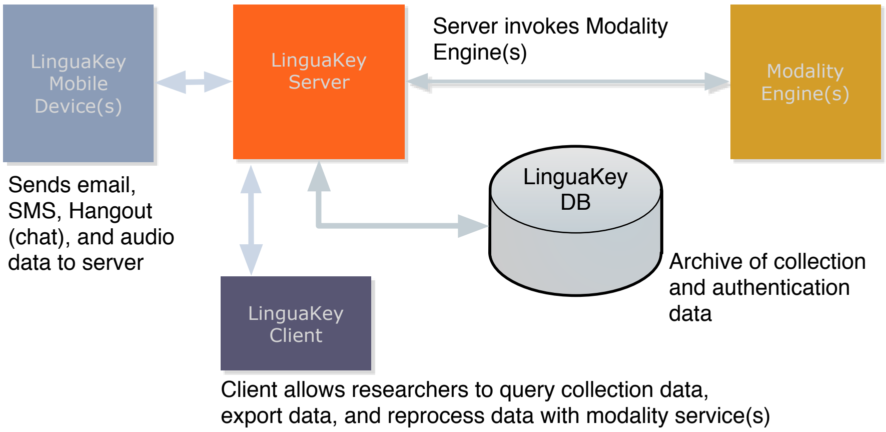
The LinguaKey process begins with the capture of speech and text on the mobile device (see Figure 3 below). This will be accomplished through tools that provide hooks into the Android operating system to record keystrokes and the microphone on the mobile device. Three initial features are extracted from speech: 1) the words being spoken are identified using a speech-to-text tool, 2) the intonation contours and prosody are identified and 3) the spectral features are extracted from every frame. For text, the input to the keystroke logger is collected, both in its raw form and the final sent message. This text then feeds a module that identifies behavioral features from the text dealing with spelling and typography. In an analogous speech behavior module, the spectral and prosodic features are modeled as a short duration biometric. The cognitive modeling component receives words from both the spoken and text inputs, as well as intonation, pitch etc. and identifies linguistic features that encompass spoken and written language use. The three modeling components output scores as log-likelihood ratios to the fusion and calibration engine which is continuously using current and recent past information to determine the probability that the current user is the authorized user.

Figure 3: LinguaKey Feature Extraction and Modeling Process



The system architecture to support this processing relies on a client/server configuration, in which the algorithms to process the data, extract features and score reside on a remote server that communicates via wireless or cellular connection. In figure 4, below the basic flow of Linguakey is depicted, with a *database* serving to contain audio and speech organized by enrolled users. A *modality* server contains models developed from different modalities of data (spoken, written, etc.). The mobile device collects data in real time from users, which is passed to the core server, which processes the data, extracts features, scores the data against the proper model and stores the audio in the database. The client allows access to the system offline for research purposes such as extracting and studying the data and running authentication experiments.

Figure 4: The LinguaKey Client/Server Architecture


LinguaKey Mobile Device(s)
LinguaKey Server
Server invokes Modality Engine(s)
Modality Engine(s)
Sends email, SMS, Hangout (chat), and audio data to server
LinguaKey DB
Archive of collection and authentication data
LinguaKey Client
Client allows researchers to query collection data, export data, and reprocess data with modality service(s)

9.4.2 Data Collection

Volunteer participant data collection is a crucial component of this effort, since data is required to understand the relationship between how language behavior and cognition manifests itself in both spoken and written modalities. Four types of data were collected in this effort over 90 sessions: 1) phone calls, 2) personal digital assistant-type queries, 3) text chat and 4) emails.

The final tally of collected data was 1,300 audio files collected totaling 1200 minutes of speech, which amounted to about 15 minutes per user. On the text side, 1,800 lines of text and 21,000 words were collected and 18,000 corrections were recorded. We are calling this collection the SpOntaneous Mobile Language Use Corpus (SOMLUC). The eventual goal is to release this data and make this widely available to the research community.

9.4.3 Research and Results

We have implemented a large set of 102 features covering everything from basic behavioral traits (punctuation, capitalization, word length, etc.) to higher level features focused on word semantics, emotional content, and language frames. Speech specific features include pitch, intonation, speech rate, and cepstrum; text specific features focus on typographic and orthographic information and corrections. We have also implemented a set of language general features that are the core of this effort, including words, speech act frames, semantics, and socio-linguistic trends.

We are currently working to understand and take advantage of entailment –the fact that features are correlated across individuals because they are associated with the same set of personality and background traits. Part of our research program is to look at how features interact to predict higher-level aspects of a person's cognitive and behavioral traits, traits that effectively characterize users but about which they have little control.

Since our goal is to combine our 100 or so features into groups that correlate with higher-level personality, cognition or demographic factors, understanding entailment within and across speech and writing is crucial.

Figure 5: Probability of starting a sentence with a lowercase letter in the SOMLUC corpus

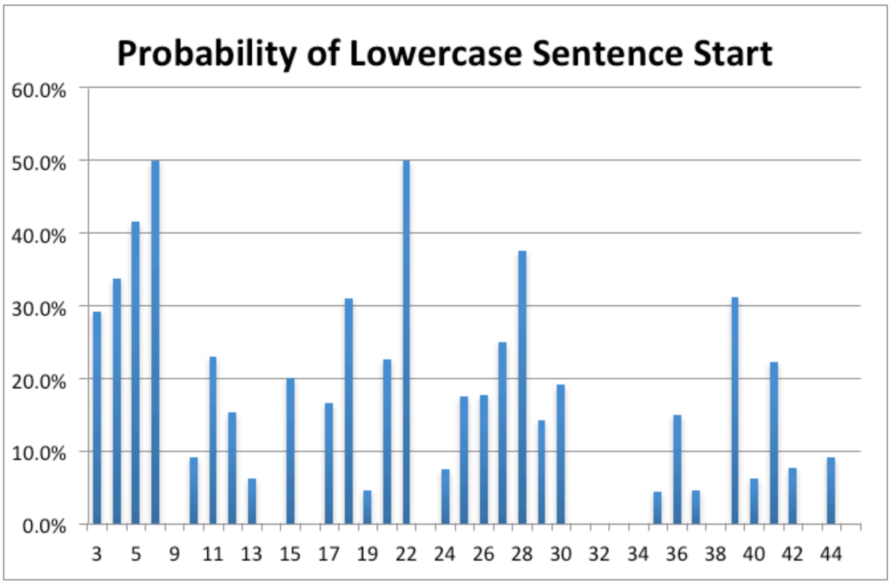
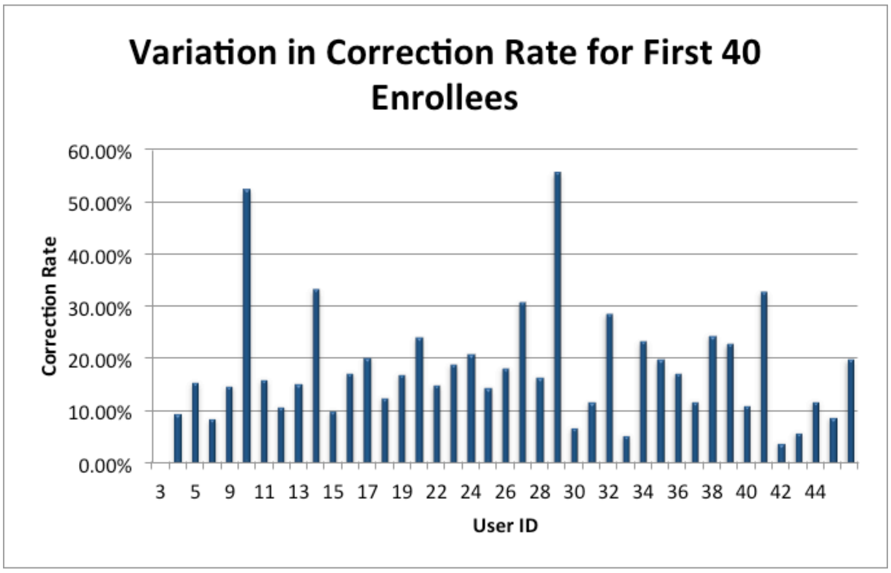


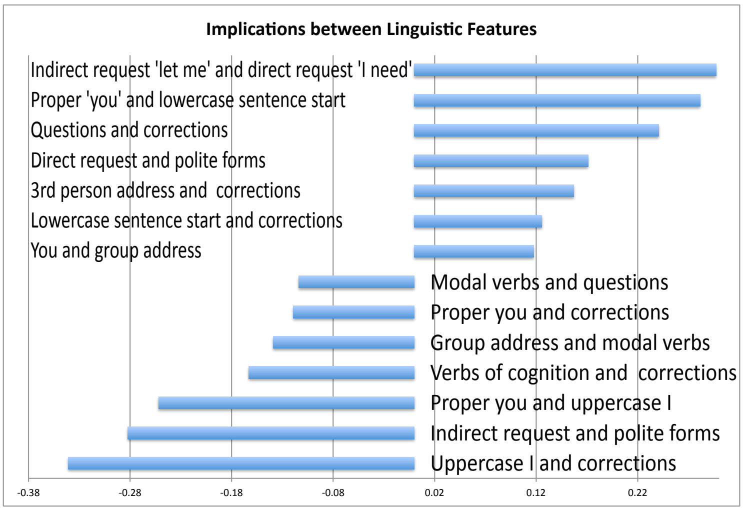
Figure 6: Differences in correction rate among users in the SOMLUC corpus



In tables 5 and 6 we present data from SOMLUC showing the variability between participants in terms beginning a sentence with a capital letter and probability that they will make a written correction. On their own, these two pieces of information are useful in terms of both characterizing individuals, and certainly correlate with aspects of an individual's personality in that careful users will tend to make corrections and use standard written forms, as do older users and more educated users based on evidence from the VERUS program.

A major focus on the LinguaKey program is to understand not just which features are useful, but to understand the relationship between features in both spontaneous writing and speech. This is important in its own right in terms of basic research as a means of clarifying the relationship between demographics and personality and types of linguistic phenomena. It is also important for the goal of authentication of users, since one of the main problems encountered in supervised system is availability of sufficient training data both in terms of raw amount and in terms of coverage of phenomena important for characterizing the user in the feature space.

Figure 7: Positive and inverse correlations between features in the SOMLUC corpus



Thus being able to predict absent features from the presence of other features is of high value in modeling an individual when limitless enrollment data is not available. In support of this approach table 7 shows the extent to which features interact and bundle, revealing higher level correlations between the phenomena and allowing us to group features into higher level classes that predict other features. For example, users who are careful to use the proper forms of “I” and “you” (rather than “i” and “u”) are naturally going to have a higher rate of corrections, since the form of the text is important to them. Likewise, users who fail to capitalize sentences are inversely correlated with correction rate, with the same rationale.

Next steps for LinguaKey will focus on the relationship between features in the spoken language, such as up-talk or filled pauses, and sociolinguistic factors in the written spontaneous language use, such as use of emoticons, questions and other phenomena associated with hedging. For example, in the VERUS data it was found that females tended to use emoticons much more than males, and that emoticons played the role of “softening” or hedging the utterance they were associated with. Likewise, “up-talk” (the persistent rising of pitch at the end of statements –Ching, 1982) is considered to be a kind of hedging phenomenon. Our hypothesis is that individuals who have pervasive up-talk in their speech will also tend to use written hedging phenomena (such as emoticons and modal verbs) more frequently. These findings will allow us to move from a set of isolated and uncorrelated features to the capability to represent the feature space as a natural gestalt, where users can be characterized effectively in a predictive space based on naturally clustering of socio-linguistically related features.

9.5 Conclusion

The future of identity detection lies in the intersection of higher-level features and machine learning –both in speech, where prosody and phonetic content are beginning to play a crucial role and writing where sociolinguistic factors are having a significant impact. In speaker identification the use of highly accurate neural network-based phone identification systems at the senone level (Lei et al. 2014) are being combined with i-vector modeling approaches to help eliminate the influence of phonetic content on speaker traits. In text processing the VERUS research demonstrated how the findings of sociolinguistics could be applied to a completely new domain and form the basis for an effective system to extract demographic information from spontaneous text chat in the virtual world environment. On-going work targets the intersection of text, speech and identity to identify the commonalities and correlations between features across language use, and factors that reflect the background, behavior, cognition and physical makeup of the individual.

References

1. N. Brümmer, FoCal II: Toolkit for Calibration of Multi-Class Recognition Scores. Software available at http://www.dsp.sun.ac.za/~nbrummer/focal/index.htm. August 2006.

2. N. Brümmer and D. van Leeuwen. On calibration of language recognition scores, in *Proceedings of the Speaker and Language Recognition Workshop,* Puerto Rico, USA, Odyssey 2006.
3. N. Brümmer et al. Fusion of Heterogeneous Speaker Recognition Systems in the STBU Submission for the NIST Speaker Recognition Evaluation 2006. In IEEE Transactions on Audio, Speech and Language Processing, 2007.
4. M. Ching. The question intonation in assertions. *American Speech 57*: 95–107, 1982.
5. N. Dehak et al. Frontend Factor Analysis for Speaker Verification. *IEEE Trans. ASLP,* vol.19, 2010.
6. E. Dieterle and J. Murray. Virtual Environment Real User Study: Design and Methodological Considerations and Implications. *Journal of Applied Learning Technology, Vol. 1, No. 1*, 19-25, 2011.
7. L. Ferrer et al. A Unified Approach for Audio Characterization and Its Application to Speaker Recognition. In *Proceedings of the Speaker and Language Recognition Workshop, Odyssey 2010,* Brno, Czech Republic, 2010.
8. L. Ferrer, L. et al. Promoting Robustness for Speaker Modeling in the Community: The PRISM Evaluation Set. In *Proc. of SRE11 Analysis Workshop,* 2011.
9. P. Flach and N. Lachiche. Confirmation-Guided Discovery of First-Order Rules with Tertius, *Machine Learning*, Vol. 42, 1/2, 2001, pp. 61-95, 2001.
10. D. Garcia-Romero. and C. Espy-Wilson. Analysis of i-vector length normalization in speaker recognition systems. In *Proc. Interspeech*, Florence, Italy, 2011.
11. M. Hall et al. 2009. The WEKA Data Mining Software: An Update. *SIGKDD Explorations*, Volume 11, Issue 1, 2009.
12. S. Herring and J. Paolillo. Gender and Genre Variation in Weblogs. *Journal of Sociolinguistics, 10(4),* 439–459, 2006.
13. S. Herring. Gender Differences in Computer-Mediated Communication: Bringing Familiar Baggage to the New Frontier. *American Library Association Annual Convention*, Miami, FL, 1994.

14. P. Kenny. Bayesian Speaker Verification with Heavy-Tailed Priors. In *Odyssey 2010—The Speaker and Language Recognition Workshop. IEEE,* 2010.
15. P. Kenny et al. A Study of Inter-Speaker Variability in Speaker Verification. *IEEE Trans. ASLP,* vol. 16, 2008.
16. C. Kim and R. Stern. Feature Extraction for Robust Speech Recognition Based on Maximizing the Sharpness of the Power Distribution and on Power Flooring. In *Proc. IEEE ICASSP,* 2010, pp. 4574–4577, 2010.
17. C. Kim. and R. Stern. Power-Normalized Cepstral Coefficients (PNCC) for Robust Speech Recognition. In *Proc. IEEE International Conference Acoustics, Speech and Signal Processing (ICASSP)*, 2012.
18. M. Kockmann et al. iVector Fusion of Prosodic and Cepstral Features for Speaker Verification. In *Proc. Interspeech,* Florence, Italy, 2011.
19. R. Lakoff. *Language and Woman's Place*. New York: Harper & Row, 1975.
20. A. Lawson et al. Sociolinguistic Factors and Gender Mapping across Real and Virtual World Cultures. *2nd International Conference on Cross-Cultural Decision Making*, San Francisco, CA, July 2012.
21. A. Lawson and N. Taylor. The Names People Play: Exploring MMOG Players' Avatar Naming Conventions. *Canadian Games Studies Association Symposium*, May 2012.
22. A. Lawson and J. Murray. Identifying User Demographic Traits through Virtual-World Language Use. *Predicting Real World Behaviors from Virtual World Data*. Ahmad, M.A., Shen, C., Srivastava, J., Contractor, N. (Eds.) London: Springer, 2014.
23. B. Lee. and D. Ellis. Noise Robust Pitch Tracking by Subband Autocorrelation Classication. In *Proc. Interspeech*, Portland, USA, 2012.
24. Y. Lei et al. Towards Noise-Robust Speaker Recognition Using Probabilistic Linear Discriminant Analysis. In *Proc. IEEE International Conference Acoustics, Speech and Signal Processing (ICASSP)*, 2012.

25. Y. Lei et al. A novel scheme for speaker recognition using a phonetically-aware deep neural network. *ICASSP 2014*, Florence, Italy, 2014.
26. A. Martin et al. The DET Curve in Assessment of Detection Task Performance. In *Proc. Eurospeech,* 1997, pp. 1899-1903, 1997.
27. M. McLaren et al. Improving Speaker Identification Robustness to Highly Channel-Degraded Speech through Multiple System Fusion. In *Proc. ICASSP*, 2013a.
28. M. McLaren et al. Improving Robustness to Compressed Speech in Speaker Recognition. In *Proc. Interspeech,* 2013b.
29. V. Mitra et al. Normalized Amplitude Modulation Features for Large Vocabulary Noise-Robust Speech Recognition. In *Proc. IEEE ICASSP,* 2012.
30. J. Murray et al. *Virtual Environment Real User Study (VERUS): Final Project Report*. AFRL-RY-WP-TR-2012-0286, Air Force Research Laboratory, 2012.
31. "NIST SRE12 Evaluation Plan," http://www.nist.gov/itl/iad/mig/upload/NIST SRE12 evalplan-v17-r1.pdf
32. J. Ohala, L. Hinton and J. Nichols. *Sound Symbolism*. New York: Cambridge University Press, 1994.
33. J. Pennebaker, R. Booth, and M. Francis. *Linguistic Inquiry and Word Count: LIWC2007 – Operator's manual*. Austin, TX: LIWC.net, 2007.
34. S. Prince. Probabilistic Linear Discriminant Analysis for Inferences about Identity. In *ICCV-11th. IEEE,* pp. 1–8, 2007.
35. S. Sadjadi and J. Hansen. Hilbert Envelope-Based Features for Robust Speaker Identification under Reverberant Mismatched Conditions. In *Proc. IEEE International Conference Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5448–5451, 2011.
36. D. Tannen. *Gender and Discourse*. NY & Oxford: Oxford University Press, 1994.
37. D. Tannen. *Conversational Style: Analyzing Talk among Friends*. Norwood, NJ: Ablex, 1984.

38. K. Walker and S. Strassel. The RATS Radio Traffic Collection System. In *Odyssey 2012—The Speaker and Language Recognition Workshop*, 2012.
39. W. Wang, W. Automatic Detection of Speaker Attributes Based on Utterance Text. *Interspeech*, Florence, Italy, October 2011.
40. C. Whissell. Using The Revised Dictionary of Affect in Language to Quantify the Emotional Undertones of Samples of Natural Language. *Psychological Reports: Volume 105, Issue 1*, pp. 509-521, 2009.