The Many Roles of Speaker Classification in Speaker Verification and Identification

Judith Markowitz

J. Markowitz, Consultants
5801 N. Sheridan Rd, #19A
Chicago, IL 60660
judith@jmarkowitz.com

Abstract. Speaker classification is a fundamental component of speaker identification and verification (SIV) technologies. This paper provides and overview of the many guises that classification takes within SIV systems.

Keywords: biometric, speaker identification, speaker authentication, speaker verification, authentication, identification, verification, speaker classification, SIV, anti-speaker, disguised voice, speaker segmentation, speaker clustering, speaker variability.

1 Introduction

One of the most widely-deployed application domains of speaker classification is within systems that perform automated speaker identification and verification (SIV). The purpose of a speaker-verification (SV) system is to determine whether the speaker is making a true or a false claim of identity. The object of speaker identification (SI) is to attach a speaker identity to a sample of speech from a previously unknown speaker. The use of both technologies is growing for security, forensics, and intelligence (Markowitz, 2000 [1], 2006 [2]).

The aim of both SV and SI is to link a speech sample to a specific individual, which is not classification. Yet, SI and SV systems (and other biometric verification and identification systems) perform a number of classification tasks in order to accomplish their goals.

2 Variability

The reason classification is used is that the data in SIV/biometric samples are variable. In fact, spoken utterances are like unique creations produced by similarities and differences arising from both external sources and the speaker. Variability is such an inherent part of SIV and other biometrics that if a sample is found to be a perfect or near-perfect match with the enrollment data from the claimed identity the system sounds a "replay" or "spoofing" attack alarm

C. Müller (Ed.): Speaker Classification I, LNAI 4343, pp. 218–225, 2007.

(Markowitz, 2005 [3]). In replay/spoofing attacks an imposter attempts to fool the biometric security system by re-using a sample taken from the claimed identity. Replay/spoofing in SV generally employs a tape recording (called a "tape attack")[1].

Resolution of variability involves classification of the speakers acoustic patterns as well as classification operations related to the communication environment (noise, device/handset type, and channel). Intra-speaker variability can be produced by speaking at different speeds, by stress, illness, fatigue, whispering; or simply by positioning the articulators (lips, teeth, or tongue) differently.

SIV systems capture and encode some intra-speaker variability during enrollment by asking for several utterances or by having the enrollee talk for up to thirty seconds while the system captures and analyzes the speech. The enrollment data are clustered into a "codebook" that describes the enrollee's voice. This information is stored as the enrollee's voice model (sometimes called "voiceprint"). It is, essentially, a delineation of the class of vocal behaviors of the enrollee.

When a new utterance is submitted to an SIV system by someone claiming to be the enrollee, the system compares the codebook for that utterance with the codebook(s) for one or more stored voice models. This process is often called the "classification" step of SIV. SV, for example, evaluates whether and how well the new sample fits into the class of acoustic patterns defined by the voice model of the person the speaker claims to be.

The most widely-used approaches for accomplishing this classification task are nearest neighbor, vector quantization, neural networks, and binary trees. Each of these techniques calculates the similarities and differences between the new sample and other voice models for each of the features utilized by the system. This process is consolidated into an overall similarity score. SV uses the score determine whether the speaker's claim of identity will be accepted or rejected; SI uses the score to rank speaker candidates for the speech sample under analysis.

Philips Speech Recognition Systems employs a variant of this technique in its speech-recognition (SR) dictation product for physicians. SR dictation systems create a separate user model for each speaker and continually update that model as the person speaks. Philips noticed that physicians often hand dictation off to assistants who use the physicians user model to do their work. If the acoustic patterns of the assistants were incorporated into the model it would degrade accuracy. The classification metric determines whether or not the current speaker is the enrolled physician. If not, it will not update the user model.

SV systems also employ a set of techniques for enhancing the accuracy of the classification called anti-speaker modeling.

[1] A human mimic could also be used to spoof an SV system but this is rare and much trickier. SIV systems employ features that reflect the size and shape of the vocal apparatus (throat, mouth, and nose) In order to mount a viable attack, the mimic must have physiology that is similar to the claimed identity or the system will detect differences.

3 Anti-speaker Modeling

Virtually all commercial and research SV systems employ some form of anti-speaker modeling. Anti-speaker modeling is designed to enhance the accuracy of an SV system by comparing the claimant's speech with voice models from speakers other than the model for the claimed identity. These additional evaluations allow the SV system to perform better in "adverse" environments, such as those with a great deal of background or channel noise, or when there is a mismatch between the handset or channel used for enrollment and that used by the claimant.

One kind of anti-speaker modeling, discriminant training, entails categorization of a newly-enrolled voice model based on comparison with all the other voices in the system. This approach is an inherent part of how neural networks and, to some extent, binary trees operate.

Another widely-used type of anti-speaker modeling is the "world model" (also called "background model"). It is a class model that is derived from the speech of a diverse population of speakers. Well-designed world models contain a balance of voices that would be representative of the voices of potential imposters.

In the world-model approach, the claimant's speech is compared with the voice model of the claimed identity and with the world model. The score is computed as a ratio of the divergence of the claimant's speech from the model of the claimed identity over the divergence of the claimant's speech from the world model (Equation 1).

$$\text{score} = \frac{\text{claimed identity}}{\text{world model}} \tag{1}$$

A high score indicates that the claimant's speech is more akin to the voice of the claimed identity than it is to the world model and that there is a high probability that the claimant is who she/he claims to be. A low score suggests that the claimant is likely an impostor.

From the perspective of speaker classification, the most interesting variant of anti-speaker modeling is cohort normalization (Higgins et al, 1991 [4]). Cohort normalization is performed when an individual enrolls in an SV system. After creating the codebook for the enrollee, the system examines its database for voice models that are similar to the newly-created model. The cohort class differs for each enrollee.

When a claimant supplies speech data to an SV system with cohort normalization the system retrieves the voice model for the claimed identity and the voice model for each of its cohorts. The claimant's speech is compared to all of those models with the expectation that, if the claimant is making a valid claim, the score for the claimed identity model will be higher than the scores for anyone in the cohort class.

The IDIAP (Dalle Molle Institute for Perceptual Artificial Intelligence), a research institute in Switzerland, employed a combination of a world model of English speakers, Arabic-speaking cohort models, and numerous examples of Osama bin Laden's speech to determine whether the 2002 recording attributed to

bin Laden was faked. Figure 1 shows that what this procedure does is determine whether or not a given sample can be categorized as being within the bin Laden class.

IDIAP [5] concluded that, "While this study does not permit us to draw any definite (statistically significant) conclusions, it nonetheless shows that there is serious room for doubt" about whether the voice on the tape could be categorized as that of Osama bin Laden.

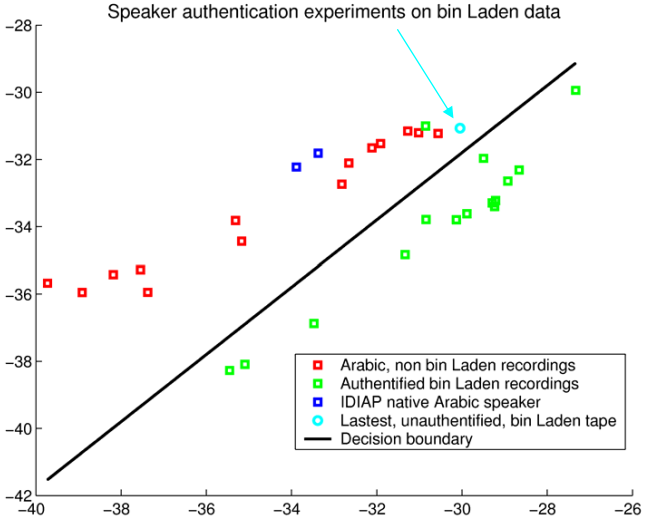

Fig. 1. 2002 IDIAP analysis of bin Laden tape [5]

4 Disguised Voices

The identification, analysis, and reversal of voice disguise are promising areas of investigation for speaker classification that are applicable to forensics and intelligence. The most systematic study of voice disguise was done by Robert Rodman (Rodman, 1998 [6]) who positioned his research on this subject within speaker classification. Rodman partitioned disguised voices into the four categories shown in Table 1 and has been since creating a database of samples for use in the development and testing of systems for identifying, categorizing, and reversing the effects of voice disguise.

The ability to detect and reverse intentional electronic disguise will be essential for the viability of SIV in the future because sophisticated voice disguise could easily merge with the work on voice forgery (usually called "voice conversion" or "voice morphing"). Voice conversion is simply the intentional electronic alteration of vocal features and patterns into the voice of a specific individual. Perrot, et al [7] assessed the threat of voice conversion to SIV systems using data

Table 1. Kinds of voice disguise [6]

Broad taxonomy of voice disguise:	DELIBERATE	NONDELIBERATE
ELECTRONIC	Electronic scrambling, etc.	Channel distortions, etc.
NONELECTRONIC	Speaking in falsetto, etc.	Hoarseness, intoxication, etc.

from the NIST speaker recognition evaluation of 2004 and found that it could pose a serious threat to existing commercial SIV technology.

5 Stress and Lie Detection

The ability to detect stress is valuable for a broad spectrum of situations in both the pubic and private sectors. It would be critical to know, for example, whether the stress levels of key employees working in nuclear weapons facilities or as international peacekeepers are too high for them to perform their jobs. A similar metric could also apply to police officers, corporate executives, and child-care workers. Being able to determine whether a suspect, informant, or witness is telling the truth would be invaluable for law enforcement and intelligence. It is equally important for business transactions and personal relationships.

Speech is an almost universal human ability. It is, therefore, fortunate that research has shown that stress affects speech in well-defined ways (Hansen and Clements, 1987 [8]; Jameson, et al, 2005 [9]; Scherer, et al. 2002 [10]). This means that stressed and unstressed speech constitute different classes of spoken behavior and that the manifestation(s) of stress in speech could be applied to the uses enumerated above.

The dominant technique for identifying stressed speech is based on "microtremor" research done in the mid-twentieth century (Lippold, 1971 [11]). Microtremors are involuntary muscular contractions that generate low-frequency oscillations (8-12 Hz) that appear to reflect the tension within muscles and seem to be part of the communication between the muscles and the nervous system. Virtually all commercial voice stress analysis and lie-detection systems utilize this approach and subsequent testing by the Air Force Research Laboratory found that these systems can distinguish stressed from unstressed speech (Haddad, et al, 2002 [12]).

Recent research reveals that stress manifests itself in a variety of ways in a person's speech (Müller et al, 2001 [13]) and that different kinds and levels of stress affect speech in different ways (Hansen, et al 2000, [14]) which indicates that stressed speech consists of a set of classes. The NATO Research Study Group (Hansen, et al, 2000 [14]) postulated four basic stressed-speech categories based on its research with military personnel. Their categories are tied to the source of the stress: physical (e.g., vibration, pressure, acceleration, equipment/physical load), physiological (e.g., alcohol, medicines, narcotics, fatigue, illness), percep-

tual (e.g., noise, poor communication channel, a listener who is having problems understanding), and psychological (e.g., emotion, lying, workload, anxiety) and produce unique constellations of effects on speech. Within and between their categories, unique constellations of effects on speech are produced. Lombard speech, for example, is a well-documented response to noise (perceptual stress) that has the following characteristics: increased vocal effort, greater duration of words due to increased vowel length, shifts in formant locations for vowels, increased formant amplitudes, and deletion of some word-final consonants (Markowitz, 1996, [15]).

The ability to go beyond microtremors is of particular interest to developers of speech recognition and SIV products because the acoustic manifestations of stress are known to cause the performance of these systems to deteriorate (Hansen, et al, 2000 [14]; Müller et al, 2001 [13]). Work by the NATO Research Study Group on Speech (Hansen, et al, 2000 [14]), the European Union Esprit VeriVox project (Karlsson, et al, 2000 [16]), and others on developing methods for transforming knowledge about stressed speech into tools for enhancing speech recognition and SIV products is still in its infancy.

6 Speaker Segmentation and Clustering

Speaker segmentation and clustering apply to the analysis of multispeaker environments. Those environments range from two-wire telecommunications channels that encode both (or all) speakers on the same channel to transcription and/or indexing of meetings and news broadcasts. In most cases, the number and identities of speakers is generally not known beforehand.

The goal of speaker segmentation is to identify all the boundaries between the speech of different speakers in the audio signal. In order to segment, the system must first determine whether the current speaker has changed. The most primitive method of detecting that a speaker has changed is to look for silence. This is useful as an alert to the system that the speaker may change but, used by itself, it is unreliable because speakers often pause in their speech (no speaker change) or talk over each other. The most common techniques for detecting that the speaker has changed are log likelihood ratio, Bayesian information criterion, and similar distance metrics (Ajmera, et al, 2004 [17]). They measure similarity/dissimilarity between the features extracted from consecutive slices (called "windows") of the signal. These approaches may be supplemented by higher-level change detectors, such as gender, language, dialect, and even topic. Boundaries are set at points where the distance measure is sufficiently large.

The next stage, speaker clustering, aims to identify, group all of the segments uttered by the same speaker, and assign a unique label to them (e.g., male No. 10, female No. 5) which are really speaker classes. Clustering employs variants of some of the same distance measures employed for establishing boundaries between speakers (Gish, et al, 1991 [18]; Reynolds, et al, 1998 [19]).

These techniques have been incorporated into automated indexing of broadcast news (Maybury, 2000 [20]), films, speeches, meetings, telephone conver-

sations, and other multi-speaker audio sources. These systems still represent emerging technology but their utility has already been demonstrated in indexing of broadcast news transmissions and intelligence gathering.

Some of these systems offer semi-automated assistance to forensics and intelligence operations. Typically, such systems identify one or more classes of speakers that match a set of criteria. One commercially-available example is the Loquendo Voice Investigation System which can be used to monitor cellular call traffic looking for speakers classifications of special interest to its law-enforcement or intelligence agency clients.

7 Conclusion

This paper has demonstrated that speaker classification is a core component of SIV applications in the real world. The "classification" step within an SIV system represents the application of speaker classification in the core SIV engine. Anti-speaker technologies extend classification to enhancements to SIV systems based on comparison of spoke data with classes of speakers. Voice disguise is an area of research for forensics and intelligence that has already been partitioned into several major classes of disguise that are currently the object of research. Systems that detect stressed speech due to emotion, cognitive load, illness, and even lying are already being used commercially. At the same time, more refined analysis of the effects of different kinds of stressors is an active area of research that is designed to make SIV more robust to intra-speaker variability caused by stress. Classification is also a critical element of systems charged with transcribing, indexing, and otherwise analyzing multi-speaker communications.

References

1. Markowitz, J.: Voice Biometrics: Speaker Recognition in the Real World. Communications of the ACM 49(9), 66–73 (2000)
2. Markowitz, J.: Speaker Biometrics: The State of the Industry. In: Proceedings of the SpeechTEK West, San Francisco (2006)
3. Markowitz, J.: Anti-Spoofing for Voice. In: Proceedings of the Biometric Consortium, Washington (2005)
4. Higgins, A., Bahler, L., Porter, J.: Speaker Verification Using Randomized Phrase Prompting. Digital Signal Processing 1(2), 89–106 (1991)
5. IDIAP: Analysis of the latest bin laden tape (2002) `http://www.idiap.ch/press_news.php/`
6. Rodman, R.: Speaker Recognition of Disguised Voices: A Program for Research. In: Demirekler, M., Saranli, A., Altincay, H., Paoloni, A. (eds.) Proceedings of the Consortium on Speech Technology in Conjunction with the Conference on Speaker Recognition by Man and Machine: Directions for Forensic Applications, Ankara, Turkey, COST250 Publishing Arm, pp. 9–22 (1998)
7. Perrot, P., Aversano, G.: R., B., Charbit, M., Chollet, G.: Voice Forgery using ALISP: Indexation in a Client Memory. In: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2005), vol. 1, pp. 17–20 (2005)

8. Hansen, J.H.L., Clements, M.A.: Evaluation of Speech under Stress and Emotional Conditions. Journal of the Acoustic Society of America 82(1), 17–18 (1987)
9. Jameson, A., Großmann-Hutter, B., Müller, C., Wittig, F., Kiefer, J., Rummer, R.: Recognition of Psychologically Relevant Aspects of Context on the Basis of Features of Speech. In: Proceedings of the Second International Workshop on Modeling and Retrieval of Context in Conjunction with IJCAI'05, Edinburgh (2005)
10. Scherer, K.R., Grandjean, D., Johnstone, T., Klasmeyer, G., Bänziger, T.: Acoustic Correlates of Task Load and Stress. In: Proceedings of the International Conference on Spoken Language Processing (ICSLP '02), Denver (2002)
11. Lippold, O.: Physiological Tremor. Scientific American 224(3), 65–73 (1971)
12. Haddad, D., Walter, S., Ratley, R., Smith, M.: Investigation and Evaluation of Voice Stress Analysis Technology (Final Report). Technical report, United States Department of Justice (Document No.: 193832) (2002)
13. Müller, C., Großmann-Hutter, B., Jameson, A., Rummer, R., Wittig, F.: Recognizing Time Pressure and Cognitive Load on the Basis of Speech: An Experimental Study. In: Bauer, M., Gmytrasiewicz, P., Vassileva, J. (eds.) UM 2001, User Modeling: Proceedings of the Eighth International Conference, pp. 24–33. Springer, Heidelberg (2001)
14. Hansen, J.H., Swail, C., South, A.J., Moore, R.K., Steeneken, H., Cupples, J.E.A., Vloeberghs, T.C.R., Trancoso, I., Verlinde, P.: The Impact of Speech Under 'Stress' on Military Speech Technology. In: NATO PROJECT 4 REPORT AC/232/IST/TG-01 Research Study Group on Speech. NATO IST/TG-01 (2000)
15. Markowitz, J.: Using Speech Recognition. Prentice Hall, Upper Saddle River (1996)
16. Karlsson, I., Banziger, T., Dankovicov, J., Johnstone, T., Lindberg, J., Melin, H., Nolan, F., Scherer, K.: Speaker Verification with Elicited Speaking-Styles in the Verivox Project. Speech Communication 31(2), 121–129 (2000)
17. Ajmera, J., Mccowan, Iain Bourlard, H.: Robust Speaker Change Detection. IEEE Signal Processing Letters **11**(8) (2004) 649
18. Gish, H., Siu, M.H., Rohlicek, R.: Segregation of Speakers for Speech Recognition and Speaker Identification. In: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '91), Toronto, Canada, pp. 873–876 (1991)
19. Reynolds, D., Singer, E., Carlson, B., O'Leary, G., McLaughlin, J., Zissman, M.: Blind Clustering of Speech Utterances Based on Speaker and Language Characteristics. In: Proceedings of the International Conference on Speech and Language Processing (ICSLP '98), Sydney (1998)
20. Maybury, M.: News on Demand. Communications of the ACM 43(2), 32–79 (2000)