SPEECH AND AUDIO TECHNOLOGY FOR ENHANCED UNDERSTANDING OF COGNITIVE RADIO USERS AND ENVIRONMENTS[*]

Scott M. Lewandowski, Joseph P. Campbell, William M. Campbell, Clifford J. Weinstein
MIT Lincoln Laboratory, Lexington, Massachusetts 02420 USA
{scl, jpc, wcampbell, cjw}@ll.mit.edu

ABSTRACT

Software-Defined Cognitive Radios, which utilize voice as a primary input/output modality, are expected to have substantial computational resources that will be capable of supporting advanced speech and audio processing applications. Yet, there has been little published research regarding how to leverage these capabilities to enhance military mission capability by building on services such as speech information extraction or background noise suppression. Such capabilities go beyond interaction with the intended user of the SDR – they extend to speech and audio applications that can be applied to information that has been extracted from voice and acoustic noise gathered from other users and entities in the environment. For example, in a military environment, situational awareness and understanding could be enhanced by processing voice and noise from both friendly and hostile forces operating in a given battlespace. In this paper, we provide a survey of a number of speech and audio-processing technologies and their potential applications to cognitive radio.

1. INTRODUCTION

Recently, there has been a significant amount of interest in applying Software-Defined Cognitive Radios (SDCR) to various military missions with the intent of increasing mission capability, effectiveness, and efficiency. Although SDCRs use the voices of the caller and callee as their primary input/output modality, there is a great deal of voice and acoustic noise information that can be gathered and exploited by the radio. Using the substantial computational resources that SDCRs are expected to have, advanced speech and audio processing techniques can be applied to the available data streams to increase the utility of the SDCR to a military user. Some of these techniques include: speaker recognition; language identification; text-to-speech; speech-to-text; machine translation; background noise suppression; adaptive speech coding; speaker characterization; and noise characterization. In this paper, we examine these technologies by: describing the technology and the current state-of-the-practice; explaining how the technology is currently being applied to or could be applied to CSDR; providing descriptions and concepts of operations for how the technology can be applied to benefit users of CSDRs; and describing relevant future research directions for both the technology and its application to CSDR. The treatment of each technology varies in its level of detail, commensurate with the availability of information and with the innovativeness and utility of the technology

2. SPEAKER RECOGNITION

Speaker recognition technologies enable systems to determine who is talking. This determination can then be used to provide user authentication for access control, identification of communicating parties, and personalization and adaptation of the device and its applications. Speaker recognition is imperfect and is characterized by two types of errors: miss and false alarm (FA). These systems are characterized by whether the speech they use is text-dependent (e.g., phrase prompted or pass phrases) or text-independent (e.g., conversational speech). The performance of these systems is quantified by two values: the false alarm rate and the false reject rate. Often, a combined measure is cited to provide a quick evaluation of overall system accuracy; this measure, known as the equal-error rate (EER), indicates the operating point at which the false alarm and false reject rates are equal. The state-of-the-art text-independent speaker recognition performance for conversational telephone speech of a few minutes in duration is in the range of 7-12% EER [1].

As introduced in [2], voice (alone, or in conjunction with face) biometrics are well suited to radios that already

[*] This work was sponsored by the Defense Advanced Research Projects Agency under Air Force contract F19628-00-C-0002. Opinions, interpretations, conclusions, and recommendations are those of the authors and are not necessarily endorsed by the US Government.

incorporate microphones (and, if applicable, cameras). Some biometrics lend themselves to continuous user authentication (e.g., to guard against lost or captured radios) and assessing varying levels of trust. For example, voice verification can be used to continuously authenticate a user while they are talking; this can be useful if the voice quality makes it difficult for the other party to determine a change in operators. A continuous authentication process might begin in a state of provisional trust and, over time, proceed in continued states of provisional trust and then to a trusted or untrusted state. While in a state of provisional trust, benign operations can be performed (e.g., adjusting radio volume), whereas sensitive operations (e.g., downloading an SDR waveform) would require a trusted state.

Voice, like other biometrics, can provide user conveniences, such as recalling preferences, biometric logins, and screen locks, which can also guard against compromised equipment losses (e.g., by disabling a radio that has been left behind). We generalize conventional biometrics by learning the users and recognizing their distinctive behaviors.

Future research directions for speaker recognition focus on making it more robust to mismatched channel conditions and applying high-level features that are somewhat like those used by humans [3].

3. LANGUAGE IDENTIFICATION

Language identification (LID) technologies allow systems to determine the language of the user from a list of possibilities. These technologies are typically available for languages such as English, Spanish, French, Arabic (various dialects), Russian, etc. These systems usually require about 30 seconds of speech to obtain good performance.

Methods for language recognition have traditionally been based upon phonetic transcription of different languages [4]. By discovering the relation between occurrences of phones (sounds like “ah” or “t”) in different languages (phonotactics), one can construct a statistical model of a particular language. A drawback of these approaches is that they require a speech recognition system to be developed in the target language, which in turn requires lexical labeling of a large corpus of speech and a phonetic dictionary that maps words into phonetic units in the target language.

An emerging class of recent methods for language recognition are based upon novel features [5]. These new features, shifted-delta cepstral coefficients, measure changes in spectrum over multiple 22 ms frames to better model language. These methods need only a speech corpus labeled with the language in order to achieve good results.

Current systems’ performance [5] is measured in terms of false alarm rate and target miss rate for detectors of individual languages. Typical error rates for speech from telephone environments are shown in Figure 1. This plot shows results for the languages: Arabic, English, Farsi, French, German, Hindi, Japanese, Korean, Mandarin, Spanish, Tamil, and Vietnamese. Results are shown for male (m) and female (f) speakers for 3 second, 10 second, and 30 second utterances. Equal error rates (EERs) are less than 3% for 30s of test speech.

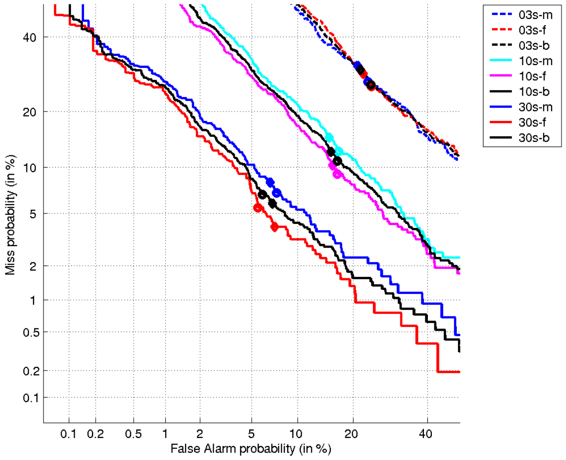

Figure 1: Typical performance of a Language recognition system. Results are taken from the 2003 NIST Language recognition evaluation.

LID has many potential applications in SDCR. First, LID could be used as a defense against system overrun; i.e., the system could allow only certain languages to be used for radio communications. A more experimental strategy may be to look for “shibboleths” to recognize the actual dialect of the speaker; e. g., does this speaker have a foreign accent? A second application of LID is in situational awareness. If speech communication can be intercepted, the language used could be determined to aid in the recognition of friends and foes.

4. TEXT-TO-SPEECH

Text-to-speech (TTS) technology automatically speaks textual information. Textual information could originate from text-based communications (e.g., e-mail, news, web, IM, chat, and SMS) or equipment display readouts (e.g., radio frequency, battery power, signal strength, network speed, time, speed, location, and bearing).

TTS could provide status information to an eyes-busy user. This would enable a warrior to focus on the mission, while hearing an explanation of their battle space and status. Different synthesized voice types (e.g., male and female) could be used to convey different types of information. For example, routine and urgent information could be conveyed in male and female voices, respectively.

The current state of TTS technology produces mostly reasonable sounding speech; however, it does not yet sound quite human. Future research directions in TTS are focusing on improving the quality of voice synthesis, pronunciation of named entities, conveyance of expression, and integration with machine translation and speech-to-text.

5. SPEECH-TO-TEXT

STT (speech to text) attempts to convert speech into a form that can be read by a user. STT includes producing entire transcripts of a conversation (continuous speech recognition), word spotting (i.e., looking for particular words), and command-and-control.

Recently, speech recognition has developed along several paths. A first path is work on large vocabulary continuous speech recognition for conversational situations. This work has been funded through projects such as DARPA EARS (Effective Affordable Reusable Speech Recognition); work in this area can be found in, e.g. [6]. Progress in STT has brought error rates down to less than 12% word error rate for telephone speech. Another recent path for STT work is in noise robustness. An overview of some of these methods can be found in [7]. Noise robustness has been studied extensively for standardization by ETSI for distributed speech recognition (DSR), as exemplified by [8]. DSR's goal is to make STT a client-server application, in which the client uses the DSR front-end to parameterize the speech while recognition is done on the server.

STT has many possible applications in SDCR. First, STT can be used for gisting – rather than having a user listen to the complete conversation, a summarized version of the output could be produced. Second, STT can be used to route certain conversations to appropriate users (see [9] and related references). Third, STT can be used for data mining speech. If radio communication is processed by STT and stored, then text-retrieval techniques (such as those used to search documents on the internet) can be a quick and efficient way of searching content. Fourth, STT can be used for command-and-control of a cognitive radio, as described in [10]. In this scenario, a speech interface frees up tactile and visual modalities so that the user can more effectively multitask. The speech interface can be used to control various aspects of the cognitive radio – radio modes, sensor interfaces, sensor analysis, etc.

6. MACHINE TRANSLATION

Machine translation (MT) automatically converts words or phrases from one language into another. This is generally done on text; however, MT can be combined with speech-to-text and/or text-to-speech to provide mixed mode translation.

MT technology could help a warrior during operations in foreign-language environments. For example, foreign-language signs, news, and radio intercepts could be roughly translated to the warrior's language to aid in understanding the battle space.

Current MT technology, as typified by various web-based systems, can be helpful for extracting some of the key words and phrases from the foreign language material, but such translations are by no means transparent, as they generally contain many errors. Transcription problems are frequent, and are often, but not always, easily detectable by users (it could be argued that it is more problematic when users are unable to detect transcription problems). Future MT related research will likely be aimed at improving basic MT performance, automatically extracting meaning, gisting, and summarization.

7. BACKGROUND NOISE SUPPRESSION

Background noise suppression is primarily used in conjunction with speech-to-text and voice communication (see Section 4 for information on the former). For the latter case, voice communication, many new technologies have become available over the last few years.

Noise suppression can be used in voice communication to enhance the effectiveness of a vocoder. In this case, a noise suppression system attempts to improve both the quality and the intelligibility of coded speech. These methods fall into several categories. First, methods that attempt to "subtract out" the noise spectrum have achieved considerable success; see, for example, [11]. Methods for spectral subtraction have been incorporated into the MELPe 1200/2400 bps update of the MELP vocoder [12]. A second class of noise suppression algorithms is based upon computational auditory scene analysis (CASA) [13]. The idea in this case is to use algorithms inspired by human processing; people effectively separate a sound field into multiple components such as music, voice, noise, etc. CASA methods use techniques such as independent component analysis and array processing to achieve noise suppression. A third class of noise suppression methods is based upon multi-modality. A well-known phenomenon for humans is that visual processing and audio processing of speech is fused (as evidenced by the McGurk effect). Several systems have tried to take effect of the visual component; see, for example, [14]. Alternate nonacoustic modalities have also been explored; these include EGG's, accelerometers, and electromagnetic sensors. Significant improvement in noise suppression has been achieved with these approaches [15]

Active noise suppression is another technology that is being incorporated into radio systems. Active noise suppression reduces the noise that a user perceives by emitting sound to cancel the undesired noise field. Active noise suppression can be used to decrease fatigue caused by exposure to high noise levels and reduce Lombard effect.

Noise suppression is a critical component of a SDCR with a speech user interface. Although not usually perceived as a cognitive capability, noise suppression is ultimately a test of a system's capability to deal with real-world conditions. Techniques such as multimodality and CASA show the sophistication and the challenge of matching human processing in this task.

8. ADAPTIVE SPEECH CODING

Adaptive speech coding is needed to fully exploit varying, limited channel capacity while achieving the goals of speech coding.[1] The current generation of speech coding standards capitalizes on the fact that people listen to speech communications systems and, thus, the systems attempt to minimize perceptual distortion. Future research will likely be focused on: optimizing coding for speech, as opposed to other types of signals; taking advantage of the language being spoken; widening the analysis bandwidths; and fusing multiple sensor streams. New adaptive speech coders will not only provide good communications-quality speech under typical conditions, but also be able to operate at dramatically reduced bit rates to conserve battery life and/or provide high processing gain to decrease the probability of a communication being intercepted and/or detected. This will yield improved and safer voice communications for the warrior.

9. SPEAKER CHARACTERIZATION

Speaker characterization is the process of determining the "state" of a user using voice processing techniques. Typically, this has meant trying to determine the emotional state that a person is in; this is typically directly related to the stress that a user is experiencing.

Speaker characterization is still a developing science. One of the difficulties is elicitation of an emotional state for corpus collection – how can an experimenter truly ensure that a participant is stressed? Another difficulty is the definition of emotional states. For example, stress can take many forms – physical stress, emotional stress, task-based stress, noise-induced stress, etc. Should all of these be separate categories of stress? Regardless of the experimental difficulties, several practical techniques for stress recognition and compensation have been examined; for examples, see the earlier work [16, 17] and work on the SUSAS corpus [18].

Speaker characterization is related to SDCR in many different ways. Speaker characterization can be part of a broader strategy of affective computing [19]. Some examples include:

- knowing the stress state of the local user as well as other users in the field to improve situational awareness;
- knowing if a user is irritated by a particular feature by relying on their voice characteristics;
- using stress level to determine appropriate modality (e.g. visual versus audio) for response to a query; and
- using verbal cues to determine if the cognitive radio made a correct decision.

Ultimately, if an SDCR is able to perceive a user's emotional state, it will make better decisions and be a more effective device.

[1] The goals include communicability, intelligible speech, quality speech, talker and state (e.g., stress) recognizibility, low delay (insignificant for push-to-talk, but must be < 300 ms total system one-way delay for normal conversation), talker and language independency, naturalness, robustness in acoustic noise (including background talkers), insensitivity to transmission errors, provide tandem (synchronous & asynchronous) coding capability, ability to transmit signaling/information tones, code at minimum rate, e.g., variable bit rate (complicates encryption, conferencing, etc.), and minimal computational and memory complexities to maximize battery life.

10. NOISE CHARACTERIZATION

Although we have considered noise a nuisance up to this point, it in fact is a useful source of information. Noise characterization can help in several areas. First, noise characterization can provide situational awareness. If a user can catalog and track the sources of noise in the environment, he can recognize anomalies that might indicate the presence of friend or foe. In this case, a noise characterization system would have to find features and provide recognition of different types of noise sources – vehicles, guns, planes, etc. Also, the directionality of noise sources would be a critical property to assess. Second, noise characterization can provide diagnostics. Noise analysis could potentially detect imminent mechanical failure of common military equipment. It could also provide a quick diagnosis of mechanical problems.

11. CONCLUSION

We have given a brief overview of several processing technologies that exploit the voice and acoustic noise streams that are likely to be available to a typical SDCR. These technologies leverage the significant computational capabilities of future SDCRs to improve the capability, effectiveness, and efficiency of military users of SDCRs. As these technologies mature and become more robust, they will provide significant force multiplication effects, which will better enable the warrior to dominate the battle spaces of the future and better function in network and information-centric warfare scenarios.

12. REFERENCES

[1] M. A. Przybocki and A. F. Martin, "NIST Speaker Recognition Evaluation Chronicles," in Proceedings of Odyssey04, 2004, pp. 15-22.

[2] J. P. Campbell, W. M. Campbell, D. A. Jones, S. M. Lewandowski, D. A. Reynolds, and C. J. Weinstein, "Biometrically Enhanced Software-Defined Radios," in Proceedings of Software Defined Radio Technical Conference, Orlando, Florida, 2003

[3] W. M. Campbell, J. P. Campbell, D. A. Reynolds, D. A. Jones, and T. R. Leek, "High-Level Speaker Verification with Support Vector Machines," in Proceedings of ICASSP, 2003, pp. 73-76.

[4] M. A. Zissman, "Comparison of four approaches to automatic language identification of telephone speech," *IEEE Trans. Speech and Audio Proc.*, vol. 4, pp. 31-44, 1996.

[5] E. Singer, Torres-Carrasquillo, P.A., Gleason, T.P., Campbell, W.M., and D. A. Reynolds, "Acoustic, Phonetic, and Discriminative Approaches to Automatic Language Recognition," in Proceedings of Eurospeech, 2003, pp. 1345-1348.

[6] R. Schwartz, Colthurst, T., Gish, H., Iyer, R., Kao, C.-L., Liu, D., Kimball, O., Makhoul, J., Matsouka, S., Nguyen, L., Noamany, M., Prasad, R., Xiang, B., Xu, D., Gauvain, J.-L., Lamel, L., Schwenk, H., Adda, G., Chen, L., and J. Ma, "Speech Recognition in Multiple Languages and Domains: The 2003 BBN/LIMS System," in Proceedings of ICASSP, 2004

[7] J.-C. Junqua, and J.-P. Haton, *Robustness in Automatic Speech Recognition*: Kluwer Academic Publishers, 1996.

[8] N. Parihar, Picone, J., "Analysis of the Aurora Large Vocabulary Evaluations," in Proceedings of Eurospeech, 2003, pp. 337-340.

[9] G. Riccardi and A. L. Gorin, "Stochastic Language Adaptation over Time and State in Natural Spoken Dialog Systems," *IEEE Trans. Speech and Audio Proc.*, vol. 8, pp. 3-10, 2000.

[10] C. Broun and W. M. Campbell, "Force XXI Land Warrior: a systems approach to speech recognition," in Proceedings of ICASSP, 2001, pp. 973-976.

[11] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Trans. Acoustics, Speech, and Signal Processing*, vol. 33, pp. 443-445, 1985.

[12] A. V. McCree, K. K. Truong, E. B. George, T. Barnwell, and V. R. Viswanathan, "A 2.4 kbit/s MELP Coder Candidate for the New U.S. Federal Standard," in Proceedings of ICASSP '96, Atlanta, Georgia, 1996

[13] M. Cooke and D. Ellis, "The auditory organization of speech and other sources in listeners and computational models," *Speech Communication*, vol. 35, pp. 141-177, 2001.

[14] X. Zhang, C. C. Broun, R. M. Mersereau, and M. A. Clements, "Automatic Speechreading with Applications to Human-Computer Interfaces," *Eurasip Journal on Applied Signal Processing*, pp. 1228-1247, 2002.

[15] T. F. Quatieri, D. Messing, K. Brady, W. M. Campbell, J. P. Campbell, M. Brandstein, C. J. Weinstein, J. D. Tardelli, and P. D. Gatewood, "Exploiting nonacoustic sensors for speech enhancement," in Proceedings of Workshop on Multimodal User Authentication, 2003, pp. 66-73.

[16] K. Cummings and M. A. Clements, "Analysis of the Glottal Excitation of Emotionally Stressed Speech," *Journal of the Acoustical Society of America*, pp. 88-99, 1995.

[17] J. Hansen and M. Clements, "Source Generation Equalization and Enhancement of Spectral Properties for Robust Speech Recognition in Noise and Stress," *IEEE Trans. of Speech and Audio Processing*, pp. 407-415, 1995.

[18] S. E. Bou-Ghazale and J. H. L. Hansen, "Speech Feature Modeling for Robust Stressed Speech

Recognition," in Proceedings of ICSLP '98, Sydney, Australia, 1998, pp. NA.
[19] R. W. Picard, *Affective Computing*: MIT Press, 2000.