Speech Adaptation in Extended Ambient Intelligence Environments

Bonnie J. Dorr, Lucian Galescu, Ian Perera, Kristy Hollingshead-Seitz, David Atkinson, Micah Clark, William Clancey, Yorick Wilks

Institute for Human and Machine Cognition, 15 SE Osceola Avenue, Ocala, FL
{bdorr, lgalescu, iperera, khollingshead, datkinson, mclark, wclancey, ywilks}@ihmc.us

Eric Fosler-Lussier

The Ohio State University, 2015 Neil Avenue, Columbus, OH
fosler@cse.ohio-state.edu

Abstract

This Blue Sky presentation focuses on a major shift toward a notion of "ambient intelligence" that transcends general applications targeted at the general population. The focus is on highly personalized agents that accommodate individual differences and changes over time. This notion of *Extended Ambient Intelligence (EAI)* concerns adaptation to a person's preferences and experiences, as well as changing capabilities, most notably in an environment where conversational engagement is central. An important step in moving this research forward is the accommodation of different degrees of cognitive capability (including speech processing) that may vary over time for a given user—whether through improvement or through deterioration. We suggest that the application of *divergence detection* to speech patterns may enable adaptation to a speaker's increasing or decreasing level of speech impairment over time. Taking an adaptive approach toward technology development in this arena may be a first step toward empowering those with special needs so that they may live with a high quality of life. It also represents an important step toward a notion of ambient intelligence that is personalized beyond what can be achieved by mass-produced, one-size-fits-all software currently in use on mobile devices.

Introduction

A major shift has been underway toward an investigation of a notion of "ambient intelligence" that transcends the "Future Mobile Revolution." The term "ambient intelligence" has generally been associated with applications that have a broad user base, as artificial intelligence "finds its way to more aspects of human life every day" (Kar, 2013). Within this paradigm, one need no more than his/her smartphone to provide information about a particular medicine, to reserve a place in the emergency room, or to request home delivery of prescription medication. Taking into account location-specific information, personalized address books, and calendar entries, the same formulas are employed for each individual who uses a given application.

However, this notion of "ambient intelligence" ignores the degree to which individuals *differ* from each other and *change* over time. What would the world be like if *each* individual were equipped with their own personal agent, designed to follow and adapt to that individual over long periods of time? Over a lifetime of co-activity, this agent would learn the user's needs, preferences, and capabilities, and would adapt accordingly. More importantly, the agent would accommodate *changes* in the user's needs, preferences, and capabilities over time. We refer to this notion of ambient intelligence as *Extended Ambient Intelligence (EAI)* below.

A major challenge in the development of EAI-based human-machine interaction (e.g., Smart Home environments designed for clients with neurological disorders) is the tradeoff between the degree of intrusiveness and presence of empathy, in addition to the impact of this tradeoff on conversational engagement (Page, 2006). When clients are under (human) observation in a hospital setting, there are frequent, intrusive interactions, yielding an environment that is less conducive to engagement and that deprives patients of perceived personal autonomy. Such an environment is associated with increased rates of mental illness and depression in older patients (Boyle, 2005). Interactions with lower levels of intrusion (e.g., remote telemedicine in the form of two-way video, email, and smart phones) may be lacking in empathy and, moreover, adaptation to clients' needs and capabilities can be lost without daily, personalized interactions. Thus, simulating empathy-based modeling of

Copyright © 2015, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

the client to guide interactions is critical (Bee et al., 2010), yet an "optimal sense of intrusiveness" needs to be maintained.

Our research at IHMC has shifted toward an extended notion of ambient intelligence in which the degree of engagement has become central in the development of systems for human-machine interaction. We have initiated efforts that focus on predicting, planning, and managing physical effects, taking into account individual behavior along psychological and social dimensions of the client, caregiver, and primary physicians (Atkinson, Clancey and Clark, 2014). We aim to explore computational models that integrate core AI components with intelligent agent architectures to enable robust, trustworthy, adaptive autonomy in the context of long-duration human-machine joint activity (Atkinson 2009; Atkinson, Friedland and Lyons 2012). Our overall research theme will be reported at the forthcoming AAAI 2015 Spring Symposium on Ambient Intelligence for Health and Cognitive Enhancement.(Atkinson, et. al., forthcoming 2015

This work necessarily involves development of AI algorithms for natural language, multi-modal social interaction, and a theory of mind.

This paper focuses on the natural language aspects of human-machine interaction, using speech recognition as an example of the need for extreme personalization and adaptation to changing conditions.

Speech Adaptation in Human-Machine Interaction

Research on EAI-based conversational agents at IHMC focuses on enabling an autonomous intelligent agent to take the role of life-long companion, providing highly personalized assistance and support for healthy assisted living. As a starting point, recent studies at IHMC with clients who have suffered Traumatic Brain Injuries (TBI) have highlighted the potential benefits of mediating communication between client and caregiver using companion agents (Wilks et al., 2014).

An important next step in the personalization and adaptation of companion-style agents is to incorporate speech recognition that accommodates different degrees of impairment that may vary over time for a given user (improvement and deterioration). Our approach involves the detection of speech language *divergences* along a range of different dimensions. We borrow the notion of *divergence* from the study of cross-linguistic variations (Dorr, 1993, 1994; Dorr et al., 2002; Habash and Dorr, 2002), where a language pair is considered "close" if it shares certain properties (but possibly not others), and "far" (i.e., divergent) if very few properties are shared.

To illustrate the concept of divergence across languages, consider three properties: vocabulary, pronunciation, and syntactic structure. Table 1 shows the properties that are shared between Spanish and four other languages. The language that diverges the most radically from Spanish is Chinese, which does not share vocabulary, syntactic structure, or pronunciation with Spanish.

Spanish	Italian	French	English	Chinese
vocabulary	X	X	X	
syntax	X	X		
pronunciation	X			

Table 1: Linguistic Divergence across Languages

We apply this same notion of divergence to the problem of "speech functioning," constraining our language pair to asymptomatic English speech compared to impaired English speech. In this case, the divergence properties to be studied are articulatory and disfluency patterns. We develop and apply techniques for detecting such divergences and leverage these to enable adaptive automatic speech recognition. The goal is to adapt to both deterioration and improvement in speech, within the same person, over time. For example, in ALS, speech is likely to become more impaired over time, whereas with TBI, the speech is likely to become less impaired. We hypothesize that while there may be variability in some static measures of impairment, there still exist trends of dynamic changes which will become clearer as we consider data over longer time spans and learn how context (patient history, environmental factors, etc.) influences or explains short-term variations in speech production.

Table 2 is a notional table illustrating a range of articulatory properties for "Baseline" English (as determined for a native English speaker who has no speech impairment i.e. pre-symptomatic) compared to those exhibited during different stages of Amyotrophic Lateral Sclerosis (ALS)—early, mid, late—thus providing a framework for capturing the degree of divergence. The final column has the highest level of divergence from the baseline: imprecise consonants, distorted vowels, and hypernasal bilabials.

"Baseline" English	ALS English (early)	ALS English (mid)	ALS English (late)
consonants	imprecise	imprecise	imprecise
vowels	~baseline	distored	distorted
bilabials (b,m)	~baseline	~baseline	hypernasal

Table 2: Example of Speech-related Divergence

Considering divergence properties as separate dimensions in speech deterioration is a crucial aspect of EAI that goes beyond standard speaker adaptation. Our work focuses not just on adaptation to a particular speaker's vocal patterns,

but also on generalization of such adaptations to other clients who are at the same stage of speech deterioration or improvement, and who therefore share common speech patterns. Furthermore, recognition of a client's level of deterioration or improvement could provide valuable data between regular visits by a caretaker or physician.

The application of the EAI paradigm to this speech problem is also an extension of our intent to consider our level of intrusiveness in the patient's life both in the end product and during data collection. While a controlled study and data collection with contextual aspects removed could be helpful for the studying divergence in those with deteriorated speech, embedding our research in an EAI environment holds the promise of learning to understand patients without requiring undue demands of their time or interfering with their daily routine.

Related Work

A cogent review of research, application, and evaluation of conversational agents that lay the groundwork for the development of today's EAI agents was published in an edited compendium (Cassell, 2000). Several of the foundational papers therein examine both verbal and nonverbal behaviors during face-to-face conversation. More recent research has focused on the development of "engagement techniques" for sustaining a conversation with a human user. Examples of such systems include SEMAINE (Schroder, 2010), VHTookit (Hartholt et al., 2013), and Companions (Wilks et al., 2011).

The closest speech processing study to the divergence approach described above is by Biadsy et al. (2011), who investigated the variation of speech properties under intoxicated and sober conditions. This earlier work was applied to the detection of intoxication (vs. sobriety), not the *degree* of intoxication. Rudzicz et al. (2014) employed another approach for recognizing impaired speech for detection of Alzheimer's (vs. no Alzheimer's) and Little (2012) developed an analogous application for detection of Parkinson's (vs. no Parkinson's) (Kepes, 2014). These approaches measure the "voice signal" to answer a yes/no question—rather than analyzing the content to determine the degree of divergence from a baseline. Nevertheless, the incidental but significant discovery from these earlier studies, that pronunciation varies systematically within categories of speech impairment, is a critical finding that can be leveraged for adapting speech recognition technology to varying degrees of impairment.

Other work has focused on finding patterns in written text which may provide evidence of a mental disorder, such as work by Rude et. al (2004). However, in such work, a person's language production ability is not impaired, but rather indicative of underlying mental factors.

Research on modeling non-native or dialectal speech (Livescu and Glass, 2000) is a closer approximation to what is suggested here for recognition of speech changes over time. The focus of dialectal modeling is on detecting divergent content, not on discerning characteristics of the speech signal. Happily, we are able to leverage an important discovery from the work of those above—as well as that of Beukelman et al. (2011), Duffy et al. (2012), Green et al. (2013), and Orimaye (2014)—which is that pronunciation varies systematically within categories of speech impairment. This discovery is critical to correlating the divergence from a baseline English and providing a foundation for adapting speech recognition technology to varying degrees of impairment.

Detection and Adaptation of Divergence

The divergences of impaired speech can be seen throughout the entire linguistic spectrum, including sub-phonetic acoustic differences, pronunciation variation, subword repetition, and disfluency in discourse. As these divergences span multiple levels of speech processing, we must develop multiple methods for identifying divergence. We anticipate using Kaldi (Povey et al. 2011) as our base speech recognition system, and plan to extend it to detect and adapt to speech divergence.

As previous work has mainly focused on detection, we will focus on the additional task of adaptation at both the phonetic and dialogue level. Traditional acoustic-phonetic adaptation techniques, such as Maximum Likelihood Linear Regression (MLLR) adaptation (Leggeter & Woodland, 1995) and Feature-space MLLR (fMLLR) (Gales 1998), seek to move either the model space or the feature representation to keep the phonetic boundaries relatively constant. We can use a traditional adaption technique as the first stage in adaptation, and then evolve it with a second transform that describes the changes over time. For pronunciation adaptation, we will incorporate the phonetic confusion networks in the pronunciation model.

In addition to detecting and adapting to sub-word disfluencies, we propose to extend techniques from our work on DialogueView (Yang et al., 2008) for annotation and detection of higher discourse-level disfluencies, including speech repairs (i.e., "like a t- sort of like a tomato") and abandoned or incomplete utterances (i.e., "what kind o-,", typically exhibited when one speaker interrupts the other).

Dialogue data from an EAI environment provides crucial information for aiding in speech repairs, by providing multi-modal context consisting of objects and environmental conditions for reference resolution and

possible substitutions. In addition, knowledge of the environment could explain away disfluencies that would otherwise be interpreted as a result of the patient's condition – we should not interpret an abandoned utterance as an indicator of speech impairment if our system detects that a pot in the kitchen is overflowing, for example.

Experimental Design

Initial data collection of impaired speech for testing will consist of de-identified speech recordings gathered from quarterly visits over 3 years of 25 ALS patients currently followed by James A. Haley Veteran's Administration Hospital (Tampa VA). We will record a mixture of context-independent calibration sentences and context-dependent conversational speech regarding events in the patients' lives.

These controlled data will be essential in quantifying the various speech and biological factors that change over time so that we have a base to build upon and compare performance to when generalizing our models to an uncontrolled ambient environment, where we can collect finer-grained data about the patient and provide more immediate feedback to caretakers and physicians.

Given the difficulty of collecting experimental data from an EAI environment in a comfortable setting for the patient, we plan on evaluating contextual effects on the speech of a baseline participant in our planned Smart Home environment equipped with microphones, depth cameras, and other sensors to monitor the occupant and the environment. We are also investigating methods of increasing the fidelity of the smart home environment and interactive context for patients through the use of immersive virtual reality and augmented reality. We will build on the results of previous studies showing that social norms and behavior carry over to virtual environments with sufficient fidelity for socio-cognitive research studies (Yee 2007; Schultze 2010). We can then combine our findings from these contextual and virtual models to determine how speech divergences can be tracked independently of context in a natural environment without interfering with the patient's daily life.

Discussion

In the endeavor to move past the current notion of "ambient intelligence," we consider an *Extended Ambient Intelligence* (EAI) paradigm that takes into account the degree to which individuals *differ* from each other and *change* over time. We consider the possibility of equipping individuals with their own personal agent, designed to follow and adapt to an individual's capabilities over long periods of time.

Within this overarching framework, we enhance human-machine interaction by providing a paradigm within which speech recognition adapts over time. There are a number of potential advantages to this adaptive view of ambient intelligence:

- We benefit from the potential for embedding this technology into several different AI systems (companions, humanoid avatar, and robot) to enable conversations with a computer.
- We leverage such paradigms to investigate interactive dialog that includes informal language understanding, in the face of disfluencies such as filled pauses (*uh*), repeated terms (*I-I-I know*), and repair terms (*she—I mean—he*).
- We are then able to investigate pragmatic interpretation of language and action: undertaking intention recognition, e.g., *Fill it with rockbee* may be understood with gesture toward a coffee cup as *Fill it with coffee.*

EAI research is well on its way toward providing a foundation for highly personalized conversational agents, e.g., in an ambient assisted living environment. Taking an adaptive approach toward technology development in this arena may be a first step toward empowering those with special needs so that they may remain in their homes and participate in society with a positive quality of life.

Acknowledgements

This material is based upon work supported by the Defense Advanced Research Projects Agency (DARPA) under Contract nos FA8750-12-2-0348 and FA8750-12-2-0348-2, by the Department of the Navy, Office of Naval Research (ONR) under award nos. N00014-13-1-0225 and N00014-12-1-0547 and by the Nuance Foundation. Any opinion, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of DARPA, ONR, or Nuance.

References

Atkinson, D.J. 2009. Robust Human-Machine Problem Solving. Final Report FA2386-09-I-4005. Air Force Office of Scientific Research. DOI: 10.13140/2.1.3923.6482

Atkinson, D.; Friedland, P.; and Lyons, J. 2012. Human-Machine Trust for Robust Autonomous Systems. In Proc. of the 4th IEEE Workshop on Human-Agent-Robot Teamwork. In conjunction with the 7th ACM/IEEE International Conference on Human- Robot Interaction (HRI 2012) Boston, USA.

Atkinson, D. J., Clancey, W. J. and Clark, M. H. 2014. Shared Awareness, Autonomy and Trust in Human-Robot Teamwork. *Artificial Intelligence and Human-Computer Interaction: Papers from the 2014 AAAI Spring Symposium on.* Technical Report FS-14-01 Menlo Park, CA: AAAI Press.

Atkinson, D.J., Dorr, B.J., Clark, M.H., Clancey, W.J. and Wilks, Y. Forthcoming in 2015. Ambient Personal Environment Experiment (APEX): A Cyber-Human Prosthetic for Mental, Physical and Age-Related Disabilities. Papers from the 2015 Spring Symposium on Ambient Intelligence for Health and Cognitive Enhancement. Technical Report SS-15-01. Menlo Park, CA: AAAI Press

Bee, N, Andre, E., Vogt, T. and Gebhard, P. 2010. The Use of Affective and Attentive Cues in an Empathic Computer-Based Companion. In Yorick Wilks (ed.), *Close Engagements with Artificial Companions: Key, Social, Psychological, Ethical and Design Issues,"* Benjamins Publishing Company:131-142.

Beukelman, D., Fager, S., and Nordness, A. 2011. Communication Support for People with ALS. *Neurology Research International.*.

Biadsy, F., Wang, W.Y., Rosenberg, A. 2011. Intoxication Detection using Phonetic, Phonotactic and Prosodic Cues. *Proceedings of Interspeech*.

Cassell, J. S., and Prevost, S. eds. 1993. *Embodied Conversational Agents*, Cambridge, MA: MIT Press, 2000.

Dorr, B. J. 1993. *Machine Translation*: A View from the Lexicon, MIT Press, Cambridge, MA.

Dorr, B. J. 1994. Machine Translation Divergences: A Formal Description and Proposed Solution. *Computational Linguistics*, 20(4):597-633.

Dorr, B. J., Pearl, L., Hwa, R., and Habash, N. 2002. DUSTer: A Method for Unraveling Cross-Language Divergences for Statistical Word-Level Alignment. *Proceedings of the Fifth Conference of the Association for Machine Translation in the Americas*, AMTA-2002, Tiburon, CA:31-43.

Duffy, J. 2012. *Motor Speech Disorders: Substrates*, Differential Diagnosis, and Management, 3rd edition.

Gales, M. J. F. 1998. "Maximum likelihood linear transformations for HMM-based speech recognition." *Computer Speech & Language* 12(2):75-98.

Green, J. R., Yunusova, Y., Kuruvilla, M. S., Wang, J., Pattee, G. L., Synhorst, L., Zinman, L., Berry, J. D. 2013. Bulbar and speech motor assessment in ALS: Challenges and future directions. *Amyotrophic Lateral Sclerosis and Frontotemporal Degeneration*. 14: 494–500.

Habash, N. and Dorr, B. J. 2002. Handling Translation Divergences: Combining Statistical and Symbolic Techniques in Generation-Heavy Machine Translation. In *Proceedings of the Fifth Conference of the Association for Machine Translation in the Americas*, AMTA-2002,Tiburon, CA: 84-93.

Hartholt, A., Traum, D., Marsella, S. C., Shapiro, A., Stratou, G., Leuski, A., Morency, L.-P., and Gratch, J. 2013. All Together Now: Introducing the Virtual Human Toolkit. In *International Conference on Intelligent Virtual Humans*, Edinburgh, UK.

Kar, S. 2013. Ambient Intelligence: Sensing a Future Mobile Revolution. *Silicon ANGLE*, Mar. 7.

Kepes, B. 2014. Using Technology To Diagnose Parkinson's Disease. *Forbes Magazine,* Mar. 27.

Leggetter, C. J. and Woodland, P. C. 1995. Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models. *Computer Speech & Language* 9(2):171-185.

Little, M. 2012. A Test for Parkinson's with a Phone Call. *TED Talk*.

Livescu, K. and Glass, J. 2000. Lexical Modeling of Non-Native Speech for Automatic Speech Recognition. In *Proceedings of ICASSP*: 1683-1686.

Orimaye, S. O., Wong, J.S., Golden, K.J. 2014. Learning Predictive Linguistic Features for Alzheimer's Disease and Related Dementias Using Verbal Utterances. In *Proceedings of the ACL Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*: 78–87.

Page, M. J. 2006. Methods of observation in mental health inpatient units. *Nursing Times*. 102(22). May 30.

Povey, D., Ghoshal, A. Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., Silovsky, J., Stemmer, G., and Vesely, K. 2001. The Kaldi Speech Recognition Toolkit. Proceedings of *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding,* Big Island, HI.

Rude, S., Gortner, E., and Pennebaker, J. 2004. Language use of depressed and depression-vulnerable college students. *Cognition and Emotion*, 18(8): 1121-1133.

Rudzicz, F., Chan-Currie, L., Danks, A., Mehta, T., Zhao, S. 2014. Automatically identifying trouble-indicating speech behaviors in Alzheimer's disease. In *Proceedings of ACM ASSETS*, Rochester NY.

Schroder, M. 2010. The SEMAINE API: Towards a Standards-Based Framework for Building Emotion-Oriented Systems. *Advances in Human Computer Interaction*, 2(2).

Schultze, U. 2010. Embodiment and presence in virtual worlds: a review. *J INF TECHNOL 25*: 434-449.

Wilks, Y., Catizone, R., Worgan, S., Dingli, A., Moore, R., Field, D., and Cheng, W. 2011. A Prototype for a Conversational Companion for Reminiscing about Images. *Computer Speech and Language*. 25(2):140-157.

Wilks, Y., Jasiewicz, J., Catizone, R., Galescu, L., Martinez, K., and Rugs, D. 2014. CALONIS: An Artificial Companion within a Smart Home for the Care of Cognitively Impaired Patients. In Proceedings of the 12th International Conference on Smart Homes and Health Telematics, Denver Colorado.

Yang, F., Heeman, P. A., Hollingshead, K., and Strayer, S.E. 2008. DialogueView: annotating dialogues in multiple views with abstraction. *Natural Language Engineering* 14(1):3-32.