A General Perceptual Model for Eldercare Robots

Timothy James Becker
University of Hartford
Department of Computer Science, 200 Bloomfield Avenue, Hartford, CT 06107
tbecker@hartford.edu

Abstract

A general perceptual model is proposed for Eldercare Robot implementation that is comprised of audition functionality interconnected with a feedback-driven perceptual reasoning agent. Using multistage signal analysis to feed temporally tiered learning/recognition modules, concurrent access to sound event localization, classification, and context is realized. Patterns leading to the quantification of patient emotion/well being can be inferred using a perceptual reasoning agent. The system is prototyped using a Nao H-25 humanoid robot with an online processor running the Nao Qi SDK and the Max/MSP environment with the FTM, and GF libraries.

Introduction

As humanoid robots become more prevalent and widespread in our society, movement toward real perceptual features must be accomplished. Without basic human perceptual abilities, an eldercare robot is just a static device. As the autonomy of the eldercare robot is increased to greater and greater levels, the need for real perceptual ability looms. The higher the level of perception that can be achieved, the more likely the patient will be to accept and trust the robot (Kanda, Ishiguro and Ishida 2001).

Researchers have been active and quite successful with specific areas of vision and audition. This research has failed however to see the overlap that exists across applications and methodologies. A general perceptual model would facilitate simultaneous access to perceptual features and permit the design of a highly autonomous eldercare robot. The combination of current research from several subfields provides a basis for the general auditory model and a framework for making best case guesses in a limited temporal window. Reinforcement learning directly elicited from the verbal mechanism of the eldercare robot combined with contextual knowledge is used to promote the mitigation of classification error and uncertainty.

Although the vision aspects of this model are not elaborated, a similar approach could be used: multi-tiered analysis that reduces information to features, recognition and classification of objects, patterns and contextual information that the perceptual reasoning agent in turn must handle.

Eldercare Interaction

While vision is the most discussed perceptual features in artificial intelligence textbooks (Russell and Norvig 2010), audition could prove to be the most useful in eldercare robot implementation (Edwards 2010). Human audition provides semantic, contextual, and emotive extraction of verbal communication and is a primary tool of the trained psychotherapist. Humans have an inherent desire to verbally communicate, and when peers are able to discern the meaning and inflection of the spoken words, they are perceived to be empathetic (although this is only true when the listening process is confirmed to be active by an emotive gestural response) (Drollinger, Comer and Warrington 2006).

In order to be effective then, an eldercare robot must be able to extract short, medium and long-term information from sound sources, while reasoning about structure and context. A system that is only trained to identify body-fall noises, spoken words or single word emotive likelihood, would miss information that arises from long-term complex sound event structures (Istrate, Binet and Cheng 2008). Ideally the eldercare robot would be able to track the speaking intonation and pacing of the patient, which in turn could infer emotional state and wellness (Edwards 2010). It is important to note however, that only by having some Natural Language Processing or NLP abilities, could automated speaking intonation and pacing lead to any of the inference mentioned. In regular speech for example, inflection often is used for questions or is used to place emphasis on important words. To create relative tracking then, normal speech must be observed within a holistic NLP view. A phoneme recognition module (Bloit and Rodet 2008) and NLP module could be generalized to provide sound classification and pattern clustering. Even with the complexity of tracking emotion and speech in mind (Vaughan et al. 2008), recognition of words and a close approximation of the speaker's f_0 or fundamental

frequency provides insightful information and has been a common element in emotion speech research (Dellaert, Polzin and Waibel 1996). Even with the successful identification of a single distress sound event (or a word that has a highly-likely distress f_0 probability), an eldercare robot would still need a memory pool to compare the existing patterns to presently emerging ones in order to determine if the spoken word was just a question, or indicative of other meaning.

To facilitate the needed patient specific adaption, reinforcement learning must be used, making the eldercare robot completely adaptive to the patient's methods of communication and sensitive to behavioral changes. Other necessary components of a complete system would be a voice synthesizer for verbal query and an emotive gesture synthesizer for displaying patient concern, neither of which will be further elaborated.

General Auditory Model (GAM)

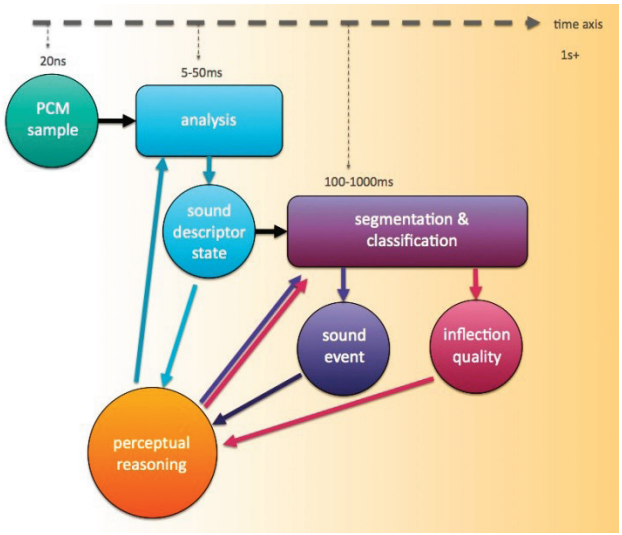

Figure 1.

A Modular General Auditory Model is proposed that is comprised of the high-level components depicted in figure 1. It is based on a preexisting musical temporal model (Jehan 2005) combined with a perceptual reasoning agent. The time axis above shows the temporal scale that governs the small and medium moving windows. Signal analysis is performed in the 5-25ms range producing sound descriptor states while segmentation and classification are performed in the 100-1000ms range producing sound events. Information that is to be retained with semi permanence is stored in the perceptual reasoning agent, which in turn uses a Data Base that can be queried by the medical overseer. Each temporal window extracts some auditory information and reduces the resolution of signal. This design not only decreases the data complexity but also minimizes the search space in each window, providing real-time response. The perceptual reasoning agent uses the recognized and reduced information to determine what events and patterns are most likely, passing a message that is indicative of its best guess.

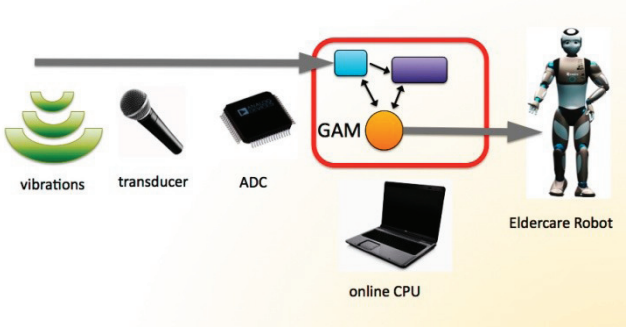

Figure 2. System Overview

Before incoming sound pressure waves can be processed in any capacity, they must first be transduced by a microphone or microphone array. For human-like sound localization, two small-diaphragm condenser microphones spaced approximately 15-17cm apart with a pinna-like baffles are needed. Computational methods exist that do not need the physical pinna-like structures (Liu et al. 2000).

Once the microphone transduces the sound pressure wave to a low-voltage and high impedance signal, it is buffered to a lower impedance line-level signal by a microphone preamplifier before the signal passes to the Analog to Digital Converter, which in turn constructs the Pulse-Code Modulated or PCM stream. The quality of the PCM stream is dependent on the bit-depth, sample rate, and precision of the internal sample clock. Proper selection of microphone, preamp, and ADC ensure that data coming in to the GAM are of the best quality (Borwick 1995), which in turn yields better overall performance.

Analysis: From Signals to Sound Descriptor States

The analysis section begins with a digital PCM stream, so the signal can be processed and transformed into meaningful information. Several contiguous samples (called a window) are collected from the PCM stream (which includes the window type, size, and overlap). The family of functions that map time domain signals to frequency domain signals are called Fast Fourier Transforms and are used in many of the algorithms that form sound descriptors (Frigo and Johnson 2005). Because a general system is desired, a two-tier analysis window is used as depicted in figure 3.

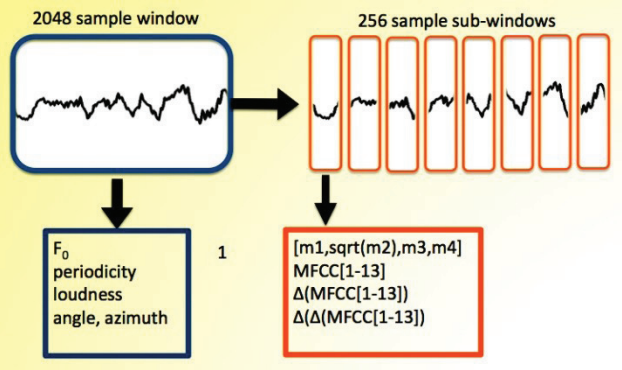

Figure 3. Two-tiered Analysis Window

In the main analysis window, fundamental frequency and periodicity are calculated with the Yin algorithm (de Cheveigné and Kawahara 2002) as well as the loudness, angle and azimuth (Liu et al. 2000).

In each of the analysis sub-windows, the spectrum moments are calculated from the power spectrum: the centroid = m1, square root of the spread = sqrt(m2), skewness = m3 and kurtosis = m4. In addition, 13 Mel-frequency cepstrum coefficients or MFCCs are computed from the Discrete Cosine Transform of the log of the Mel-bands, as well as the delta and delta-delta inter-frame difference functions. Collectively these sub-window parameters approximate the shape of the signal spectrum and changes made to it.

By setting the main window's hop size to the same value used in the sub-window's FFT, the resulting scaled vectors can be recombined into a normalized matrix called a sound descriptor which updates with every new hop. This matrix of 46 sound descriptors excluding localization angle and azimuth (which are only used for determining where the sound is emanating from) essentially provides access to all the aspects of the auditory stimulus needed for segmentation and classification.

Segmentation and Classification: From States to Events

Sound descriptor states are joined together sequentially and segmented by a threshold function (using loudness, periodicity, and the spectral centroid). This sequence can then be referred to as a sound event, which is a medium time scale unit in the range of 200-1000ms.

To classify both English phonemes and general sounds, modifications to the Short-Time Viterbi (STV) Hidden Markov Model (HMM) algorithm (Bloit & Rodet 2008) were made. This algorithm achieves near offline accuracy with online latencies that allow best guessing within the limited sliding window calculation. It would be best, given the nature of eldercare to have a large corpus of region specific aged speech (or even better to train the system with the patient's voice), but since this doesn't currently exist, a freely available American English corpus should be used instead to train the phoneme models (Pitt et al. 2007). In addition to phonemes, the training set should include all manner of environmental sounds, which could be quite a large task if the application of eldercare use was not already selected. And eldercare specific sound corpus would include a wide variety of household objects, alarms, sirens, body-fall sounds and musical excerpts. The STV sound classification method takes a 39 MFCC value sub-matrix from the sound descriptor state. A variable state size observation window is used to calculate the Viterbi path yielding a real-time guess of the maximum likelihood of the sound event (Bloit and Rodet 2008).

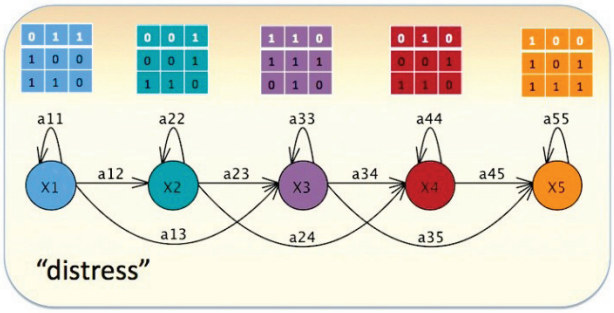

Figure 4. left-to-right HMM with state skipping.

The remaining seven features feed a left-to-right HMM (depicted in figure 4) purpose is to classify inflection quality in the voice of the patient (Bevilacqua et al. 2009). F_0, loudness, periodicity, and the spectral moments are used to train the inflection quality, although other researchers have used these features to train so called emotive states: Anger, Sadness, Happiness, Fear, Disgust, Joy and Surprise (Ververidis and Kotropoulos 2003). It is beyond the scope of this discussion to determine whether emotive states in speech are truly being measured, or whether inflection quality is a more accurate name. Assuming inflection quality is more readily observable, the enumerated approximations to Anger, Sadness, Fear and Disgust could be used to infer the presence of a distress emotive state. As with the sound recognition, a real-time guess of maximum likelihood is produced as synchronized intervals. Both of these guesses are then passed along to the perceptual reasoning agent where context and relevance is determined.

Perceptual Reasoning Agent

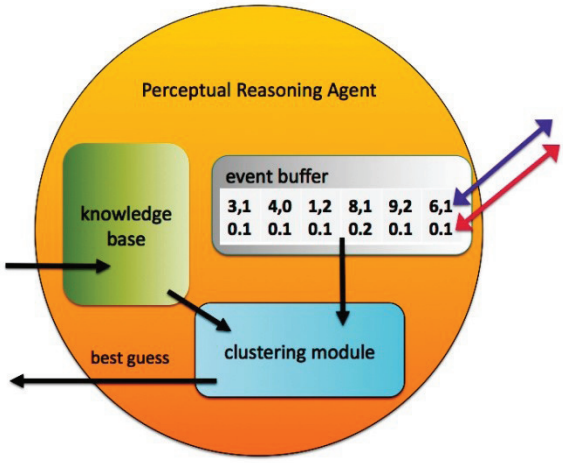

Figure 5.

The main goal of the perceptual reasoning agent is to determine if phonemes or sounds are present by applying a clustering model to the incoming event buffer as depicted in figure 5. The event buffer is a (10-60s) recirculating array that appends a temporal distance stamp to incoming guesses. An online clustering method must be used at this stage as well, incorporating elements of a language model for phoneme-to-text (Mori, Takuma and Kurata 2006) and sound event- to-sound structure tasks (Jehan 2005). Training of the clustering model occurs from a speech corpus and the normal speech from the patient. With a fully autonomous system, the patient's speech would slowly remove the presence of the original training weights, decreasing general recognition accuracy and in turn increasing the recognition of the patient's speech. Given a bounded time frame, the clustering module applies its best guess using a dictionary of words and sound patterns, giving preference to word recognition. Clustered words are output as text, while sound patterns output by a unique ID. The output in a complete system would connect to behavioral or other reasoning mechanisms. The arrow leaving the perceptual reasoning agent shown in figure 5 is used to provide reinforcement learning to specific recognized words. In the case of distress sounds, confirmation of a verbal query would do nothing but proceed, while a declination would cause the input arrow on the left to make modification to the knowledge base. This iterative technique makes long-term adaption feasible and increases the likely hood of recognizing disturbances in the patient's voice.

Prototype

A proof of concept system was constructed with an Aldebran Robotics Nao H25 Humanoid Robot (Academic Edition v3) and an online processor (Apple MacBook Pro laptop) communicating via Ethernet (for reduced latency and increased speed). Due to a flaw in the H25 design (the microphones are located next to a CPU fan), external pressure gradient condenser microphones were used to feed discrete Grace Audio Designs M101 microphone preamps and an RME Fireface800 ADC connected to the laptop. This system provided an idealized input signal which was essential for evaluating the software components and algorithms used in this tiered design.

The GAM software was constructed using MAX/MSP visual object-based design software (Puckette et al. 2011) with the addition of the IRCAM FTM, MNM, and GF libraries (Schnell et al. 2005) (Bevilacqua et al. 2009). Models were built with the Buckeye Speech Corpus (Pitt et al. 2007) as well as a corpus of household objects and environmental sounds recorded by the author using a Sony PCM-D50 field recorder. Although lowpass filtering and downsampling of the speech corpus was already conducted, full bandwidth sound recordings with a 44.1 KHz sample rate were used.

The GAM was realized as described in figure 1 with separate analysis, classification and perceptual reasoning modules. Because MAX/MSP provides objects for network communication, audio processing, vector/matrix operations and SQLite3 DB functionality, it was the ideal experimental test bed for perceptual this AI research. Word/sound pattern recognition was conducted by using KNN style clustering trained with keywords and sound patterns that were thought to be useful for eldercare use.

Messages were then passed over the network to the Nao H25 using the Nao Qi SDK, where behaviors were launched and the prebuilt text-to-speech vocal synthesizer provided patient query. Responses were used to make modification to the knowledge base when in contrary to the best guess, and confirmation was defaulted when a response was lacking within the event buffer after a reasonable time period.

Conclusion

The eldercare robot is an ideal vehicle for evaluating integrated perceptual models. The physical interaction with the environment and patient allow the audible confirmation of wellness. This confirmation process provides reliable training data, making the perceptual reasoning agent more useful with time and capable of learning in a very human-like manner (Yule 2010).

Methods in the analysis and sound classification phase have yet to be extensively compared with alternate

methods. Further evaluation of varied learning methods, parameters and inputs is needed. Preliminary results are promising and could lead towards the adoption of a more generalized auditory model in humanoid robot design. The complex work of identifying overlap in analysis methods and machine learning algorithms still remains, but immerging research in online learning offers great promise (Bifet et al. 2010) (Bevilacqua et al. 2009) (Maxwell et al. 2009) (Schuller et al. 2010).

Acknowledgements

Dr. Michael Anderson and the Department of Computer Science at the University of Hartford for guidance and financial support, respectively.

References

Bevilacqua, F., Zamborlin, B., Sypniewski, A., Schnell, N., Guédy, F., et al. 2009. Continuous realtime gesture following and recognition. *Gesture in Embodied Communication and Human-Computer Interaction*, 73–84. Athens Greece.

Bifet, A., Holmes, G., Pfahringer, B., Kranen, P., Kremer, H., Jansen, T., et al. 2010. MOA: Massive Online Analysis, a Framework for Stream Classification and Clustering. *HaCDAIS 2010*, 3.

Bloit, J., & Rodet, X. 2008. Short-time Viterbi for online HMM decoding: Evaluation on a real-time phone recognition task. *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2121-2124. Las Vegas, Nevada, USA

Borwick, J. Ed. 1995. *Sound Recording Practice*. Fourth edition, pp. 37-84). New York, New York: Oxford University Press.

Cheveigné, A. de, & Kawahara, H. 2002. YIN, a fundamental frequency estimator for speech and music. *The Journal of the Acoustical Society of America*, *111*(4), 1917.

Russell, S. and Norvig, P. eds. 2010. *Artificial Intelligence: A Modern Approach*, Third edition, p.929-970: Upper Saddle River, NJ: Prentice Hall.

Dellaert, F., Polzin, T., & Waibel, A. 1996. Recognizing emotion in speech. *Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP '96*, 1970-1973. Beijing, China

Drollinger, T., Comer, L. B., & Warrington, P. T. 2006. Development and validation of the active empathetic listening scale. *Psychology and Marketing*, *23*(2), 161–180.

Edwards, S. 2010. Definitive Theory of Human BioAcoustic Vocal Profiling & Sound Presentation. *European journal of clinical nutrition*, *64*(8), 914.

Frigo, M., & Johnson, S. G. 2005. The Design and Implementation of FFTW3. *Proceedings of the IEEE*, *93*(2), 216-231.

Istrate, D., Binet, M., & Cheng, S. 2008. Real time sound analysis for medical remote monitoring. *Proceedingsof the. Annual International Conference of the IEEE EMB Society.*, 4640-3. Vancouver, BC.

Jehan, T. (2005). *Creating Music by Listening*. *Doctoral Dissertation*. Massachuesetts Institute of Technology. Cambridge, MA.

Kanda, T., Ishiguro, H., & Ishida, T. 2001. Psychological analysis on human-robot interaction. *Proceedings 2001 ICRA. IEEE, Cat. No.01CH37164*, 4166-4173. Seoul, Korea.

Liu, C., Wheeler, B. C., O'Brien, W. D., Bilger, R. C., Lansing, C. R., & Feng, a S. (2000). Localization of multiple sound sources with two microphones. *The Journal of the Acoustical Society of America*, *108*(4), 1888-905.

Maxwell, J. B., Pasquier, P., & Eigenfeldt, A. 2009. Hierarchical sequential memory for music: A cognitive model. *International Society of Music Information Retrieval*.

Mori, S., Takuma, D., & Kurata, G. 2006. Phoneme-to-text transcription system with an infinite vocabulary. *Proceedings of the 21st ICCL and the 44th annual meeting of the ACL*, 729-736. Morristown, NJ, USA:

Pitt, M. A., Dilley, L., Johnson, K., Kiesling, S., Raymond, W., Hume, E., et al. 2007. Buckeye Corpus of Conversational Speech, 2nd release. Columbus, OH: Department of Psychology, Ohio State University.

Puckette, M., Zicarelli, D., Sussman, R., Kit Clayton, J., Berstein,, et al. 2011. MAX/MSP 5.1.8. Cycling74, IRCAM. Retrieved November 5, 2011, from cycling74.com.

Schnell, N., Borghesi, R., Schwarz, D., Bevilacqua, F., Muller, R. 2005. FTM—Complex data structures for Max. *In Proc. 2005 ICMC*. Barcelona, Spain.

Schuller, B., Metze, F., Steidl, S., Batliner, A., Eyben, F., et al.. 2010. Late Fusion of Individual Engines For Improved Recognition Of Negative Emotion In Speech. *2010 IEEE ICASSP.* p. 5230–5233. Dallas, Texas.

Vaughan, B. J., Cullen, C., Kousidis, S., & McAuley, J. 2008. Emotional speech corpus construction, annotation and distribution. *LREC 2008.* p. 5. Marrakesh, Morocco.

Ververidis, D., & Kotropoulos, C. (2003). A state of the art review on emotional speech databases. *Proc. 1st Richmedia Conference* (p. 109–119).

Yule, G. (2010). *The Study of Language* (Fourth., p. 320). New York, New York, USA: Cambridge University Press.