Speech Processing Technology and Challenges for Wearable Devices

Randy Gomez
Honda Research Institute
8-1 Honcho Wako-shi, Japan
r.gomez@jp.honda-ri.com

Kazuhiro Nakadai
Honda Research Institute
8-1 Honcho Wako-shi, Japan
nakadai@jp.honda-ri.com

Keisuke Nakamura
Honda Research Institute
8-1 Honcho Wako-shi, Japan
k.nakamura@jp.honda-ri.com

Levko Ivanchuk
University of Manitoba
Winnipeg Canada
lvanchuk@cs.umanitoba.ca

Takeshi Mizumoto
Honda Research Institute
8-1 Honcho Wako-shi, Japan
t.mizumoto@jp.honda-ri.com

Pourang Irani
University of Manitoba
Winnipeg Canada
Pourang.Irani@cs.umanitoba.ca

Yurii Vasylkiv
Honda Research Institute
8-1 Honcho Wako-shi, Japan
vasilkivyra0202@gmail.com

Abstract

In this paper we present speech technology from a signal processing point of view to enable researchers and practitioners in the areas of shared interest a different perspective of the available tools, concepts, operation assumptions and limitations of the technology. First, we introduce several of the research activities at the Honda Research Institute Japan (HRI-JP). We then present a case study of the complementary impact of multimodal processing for improving speech-related interactions. Consequently, we present some challenges in expanding our current speech technology advances to mobile and wearable devices. Lastly, we outline the roadmap of our research collaboration with the HCI laboratory at the University of Manitoba to effectively use speech technology for wearable devices.

Author Keywords

Microphone array; Speech technology; Wearable devices; Multimodal

ACM Classification Keywords

H.5.m [Information interfaces and presentation (e.g., HCI)]: Miscellaneous; See [http://acm.org/about/class/1998/]: for full list of ACM classifiers. This section is required.

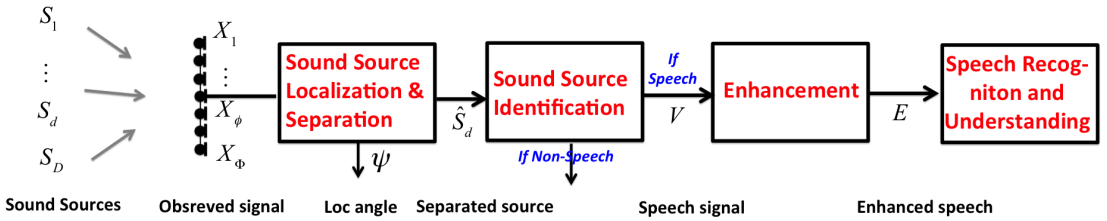


Figure 1: Block diagram of a speech processing system using microphone array.

Introduction

Speech communication is a basic form of human expression. We use speech along with other modalities whenever we interact with one another. More recently, this form of communication is being adapted for interacting with other devices, including robots. Hence, speech modality is becoming a key component in human computer interaction. Over the last couple of decades, speech technology, such as microphone array processing and speech recognition have been mostly associated with desktop computers, smart TVs and robots. This was primarily due to the resource constraints imposed by such a modality.

With advances in microchip design, resulting in nano-scale, powerful and power-efficient microprocessors, small form factor mobile and wearable devices have become pervasive. As these devices gain prominence among the ecosystem of interconnected devices, the conventional modality of speech interaction needs re-examination to suit the platform of interconnected wearable devices. Wearable devices have different design requirements than their larger form factor counterparts (e.g. robots, smart TVs, etc.) which pose numerous challenges. This leads to novel opportunities for a new paradigm and novel approaches in speech technology.

At HRI-JP, our research activities have been focused on speech technology concepts applied to robots, smart house and smart TVs. Moving onto wearable devices presents an opportunity to expand our reach. Moreover, we believe that for speech technology to maintain its impact toward the future, the challenges posed by wearable devices must be addressed. To facilitate this effective adoption of speech technology for wearable devices, HRI-JP and the University of Manitoba HCI Laboratory has embarked in a joint research collaboration. Areas of interests include speech UI design, evaluation metrics for speech input on wearable devices and applications of wearable speech-enabled devices.

Research Activity in Acoustic Processing at Honda Research Institute Japan (HRI-JP)

The block diagram shown in Figure 1 summarizes the major components of speech and audio processing (i.e., from sound source localization to speech recognition and understanding). A microphone array system suppresses unwanted spatial noise sources and recovers the desired speech source. Let $X_\phi(\omega, f)$ and $S(\omega, f)$ denote the input acoustic signal of the ϕ-th channel ($1 \leq \phi \leq \Phi$) and a sound source signal after *Short Time Fourier Transform (STFT)*, respectively. ω denotes frequency domain while f denotes the frame index. The room transfer function (TF), $\boldsymbol{A}(\omega, \psi)$ is a vector of TF for each microphone ϕ.

Sound-Source Localization

We compute a correlation matrix of $\boldsymbol{X}(\omega, f)$ and extract its eigen values. Let $\boldsymbol{E}(\omega, f) = [\boldsymbol{e}_1(\omega, f), \ldots, \boldsymbol{e}_\Phi(\omega, f)]$

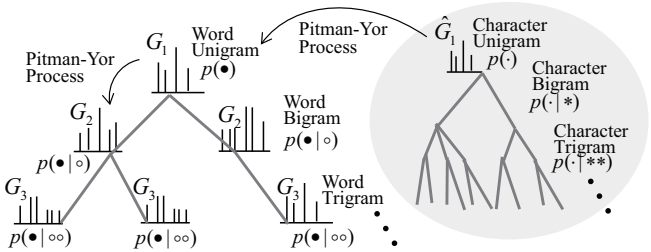


Figure 2: Word and Character N-gram Models in NPY Process

denote the eigen vectors. The spatial spectrum is given as,

$$P(\omega, \psi, f) = \frac{|\boldsymbol{A}^*(\omega, \psi)\boldsymbol{A}(\omega, \psi)|}{\sum_{\phi=L_s+1}^{\Phi} |\boldsymbol{A}^*(\omega, \psi)\boldsymbol{e}_\phi(\omega, f)|}, \quad (1)$$

where $()^*$ is a complex conjugate transpose operator, and L_s is the number of sound sources. The resulting ψ is the localization angle that maximizes Eq. (1) based on *MUltiple SIgnal Classification (MUSIC)*[9] [7].

Sound-Source Separation

We combine both the methods based on blind separation and beamforming referred to as *Geometric High-order Decorrelation-based Source Separation (GHDSS)* [8]. In this method, only the beamforming part uses TFs to overcome permutation and scaling problems. Next, we estimate a separation matrix, denoted as $\boldsymbol{\Omega}(\omega, f)$, so that $\boldsymbol{\Omega}(\omega, f)\boldsymbol{X}(\omega, f)$ converges to $S(\omega, f)$. GHDSS iteratively computes $\boldsymbol{\Omega}(\omega, f)$ to minimize cost function $J(\boldsymbol{\Omega}(\omega, f))$, which is described as:

$$J(\boldsymbol{\Omega}(\omega, f)) = \alpha J_1(\boldsymbol{\Omega}(\omega, f)) + (1-\alpha)J_2(\boldsymbol{\Omega}(\omega, f)), \quad (2)$$

where $J_1(\cdot)$ and $J_2(\cdot)$ denote cost functions for blind separation and geometric constraints, respectively. α is a weighting parameter.

When multiple spatial acoustic events exist, $\boldsymbol{X}(\omega, f)$ contains mixture of the events. We assume that there is a unique sound event in a location, and a sound event contains only one sound source. Let $S_d(\omega, f)$ denote the d-th spatial acoustic event ($1 \leq d \leq D$). To obtain $S_d(\omega, f)$, we compute $P(\omega, \psi, f)$ in Eq. (1), localize and detect the d-th acoustic event by a thresholding and tracking approach [1]. For each localized event, we conduct sound source separation using Eq. (2) to estimate $S_d(\omega, f)$, denoted by $\hat{S}_d(\omega, f)$.

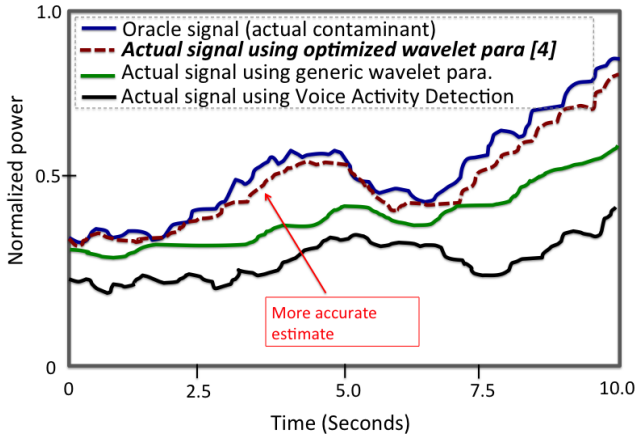
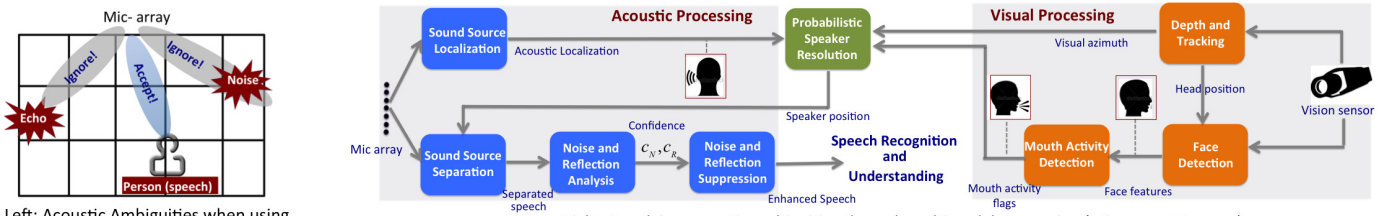


Figure 3: Power estimation of contaminant signal (noise + late reflection)

Sound-Source Identification

We employ a technique inspired by natural language processing in which the acoustic speech is defined by words and grammar. Similar to the natural language technique, the segmentation of $\hat{S}_d(\omega, f)$ is important to improve noise-robustness. Unlike speech in which segmentation is straightforward, acoustic events (non speech sounds) such prior segmentation does not exist. For a segmentation without any prior knowledge, we employ Nested Pitman Yor (NPY) [2], originally proposed for morphological analysis of natural languages. NPY process essentially maximizes the prob-



Left: Acoustic Ambiguities when using microphone array.

Right: Resolving acoustic ambiguities through multimodal processing (Mic array + Camera).

Figure 4: Improved unwanted sound rejection and speech recognition performance through multimodal-processing (mic-array + camera).

ability of the following equation for given $\hat{\mathbf{c}}_d$ with arbitrary word sequences $\mathbf{w}_d = w_1 \dots w_{N_{wd}-1} w_{N_{wd}}$:

$$\mathbf{w}_d^* = \underset{\mathbf{w}_d}{\operatorname{argmax}}\, p(\mathbf{w}_d|\hat{\mathbf{c}}_d)\ , \tag{3}$$

where N_{wd} is the number of words in $\hat{\mathbf{c}}_d$, which are estimated by the NPY process, and $\hat{\mathbf{c}}_d$ is the extracted features of $S_d(\omega, f)$. To solve Eq. (3), NPY process models languages as a combination of a word N-gram model and a character N-gram model. The two N-gram models are estimated like that shown in Figure 2. As a result, the signal $\hat{S}_d(\omega, f)$ can be classified as either speech (i.e.,$V(\omega, f) = \hat{S}_d(\omega, f)$) or any other acoustic event (non-speech). We note that source identification can also be employed using Gaussian mixture models.

Speech Enhancement

The separated speech source $V(\omega, f)$ may contain some traces of noise and reverberation referred to as contaminants. The former is additive in nature while the latter is treated as channel distortion. Reverberation is a phenomenon caused by the different time delays of the reflections of the acoustic signal as observed by the microphone sensor inside an enclosed environment. Noise and reverberation creates mismatch which degrades speech recognition performance. Speech enhancement is conducted to suppress both noise and reverberation. In our method [4][5], speech enhancement is conducted in the wavelet domain such that $V(\omega, f) \Leftrightarrow V(\upsilon, \tau)$, where υ and τ are the scaling and shifting parameters or referred to as wavelet parameters. Wavelet-based Wiener filtering is employed by weighting the contaminated separated source with the Weiner gain as

$$E(\upsilon, \tau) = V(\upsilon, \tau)\ \ .\ \ \kappa, \tag{4}$$

where $E(\upsilon, \tau)$ and κ are the enhanced signal and wiener gain respectively, where κ is expressed as

$$\kappa = \frac{V(\upsilon, \tau)^2}{V(\upsilon, \tau)^2 + R(\upsilon, \tau)^2 + B(\upsilon, \tau)^2},$$

where $V(\upsilon, \tau)^2$, $R(\upsilon, \tau)^2$ and $B(\upsilon, \tau)^2$ are the wavelet power estimates for the speech, reverberation (i.e. late reflection), and background noise, respectively. We note that the enhancement quality is dependent on κ. Hence, system performance depend on the power estimation capability of the system. By using the optimized values for υ and τ as described in [4], we can compute the respective power estimates directly from the observed contaminated signal (separated speech) effectively. Figure 3 shows the effectiveness of our method's power estimation scheme.

Automatic Speech Recognition and Understanding

The enhanced speech signal is then recognized by evaluating the likelihood $P(E|L;\lambda)$ by the ASR. Where L and λ are the language and acoustic models, respectively. The

resulting hypothesis is further processed in order to extract the meaning (understanding). Methods such as conditional random field (CRF), support vector machine (SVM), DNN, etc. are usually employed for spoken language understanding [3] [6].

3. Robust Multimodal Speech Processing

Robustness in system performance can be achieved by combining different modalities. Figure 4 (Left) shows a typical problem of acoustic ambiguities in which non speech signals (i.e., speech and echo) are observed by the microphone array and processed by the speech recognition system. This problem is addressed by incorporating both visual processing to the existing acoustic processing discussed in the previous section to spatially reject unwanted non speech sources as depicted in Figure 4 (Right). This method assumes that only an acoustic signal with a corresponding detected face constitutes a valid speech input. Moreover, mouth activity information further supplements the notion of a valid speech. By implementing Figure 4, rejection of the unwanted signals is improved, resulting in a more robust speech recognition performance.

Challenges

There are factors affecting the smooth adoption of conventional speech technology (i.e., microphone array system) to wearable devices. One of these is its small form factor. As a result, not all of the features of typical microphone arrays would be readily available to wearable devices. However, it is not yet clear whether all of the features are needed or not on such devices. One task would require investigating what aspects of a speech technology would fit common tasks and applications on such platforms. Further, the following are some of the concerns that need to be addressed:

- Sensor Quantity: To effectively resolve the number of sound-sources, the number of sensors should be at least equal to the number of sound sources ($\Phi \geq D$ in Figure 1). The limited form factor can severely impeded the number of potential sensors that can be deployed. One solution may involve affixing sensors to one or more wearable devices, to accommodate for a broad number of sources.
- Sensor spacing: The spacing of microphones affect the frequency response of the system. Sensors that are closely spaced result in a poor low frequency resolution than sensors that are placed further apart. The speech signal is concentrated at the low frequency spectrum. This means that low frequency response for a microphone array on wearable devices is not as good as those for smart TVs.
- Geometry: The geometry of the microphone array also affects the ability to spatially discriminate unwanted noise sources. There are geometric limitations on wearable devices, and engineering the locations of the microphones in the array is a significant task.
- Computation Cost: Most speech processing algorithms (i.e. microphone array processing) are computationally expensive. Cloud processing may be necessary.

Enhancing Research Through Joint Collaboration

A joint partnership has been identified to leverage the strengths of two labs: one specializing in speech technologies and the other on interfaces for mobile and wearable devices and applications. While the collaboration has only recently being unfolding, we believe several aspects will lead to a fruitful collaboration. We also seek to engage the community, through this workshop and others, as to how best to lead such collaborations to advance emerging areas.

The collaboration includes several facets. First, is the free exchange of intern students, who are able to cross-pollinate ideas from both labs to advance their own theses or research plans. This has been on-going and several others are planned in the near future. A second feature, is the identification of novel application areas, to immediately deploy such technologies in the wild. While much design and experimentation will take place in our respective labs, discussions are underway to already experiment with such technologies in specific vertical markets. In nursing homes, our goal is to devise speech-enabled wearables to monitor the activity of the elderly and to detect events such as falls, or dangerous events, such as aggression as is common in some units (Alzheimer's unit, for example. In manufacturing environments, we are interested in isolating noise from the environment, to robustly capture speech input for interaction with head-worn displays or other devices in the user's vicinity. We believe such input is key in such hands-busy and noisy environments. On general purpose consumer devices, such as smartwatches, our goal is to devise input techniques that sit at the confluence of direct input (via touch or swipes) and speech. Such forms of multimodal interfaces are subject to the many variations in user movement, environment and hardware limitations. While we have initially targeted these specific areas for initial exploration, we are not excluding others, including the use of such technologies for children or the elderly.

Conclusion

This position paper provides a brief overview of some approaches in speech technologies, as these have demonstrated success in environments and applications of value to HRI. Moving forward, we have identified a collaboration that will open new possibilities for speech input on small, wearable devices. While we have identified several immediate challenges and areas of applications, we encourage and welcome a discussion on how to move forward and make speech modality on wearables as common as it is in our daily interactions.

References

[1] K. Nakadai et. al. 2009. Design and Implementation of Robot Audition System HARK. In *Advanced Robotics*. vol. 24, pp. 739–761.

[2] D. Mochiahshi *et al.* 2009. Bayesian unsupervised word segmentation with nested Pitman-Yor language modeling. In *Proceedings of the Joint Conf. of ACL and AFNLP*. vol. 1, pp. 100–108.

[3] K. Yao *et al.* 2014a. Recurrent Conditional Random Field for Language Understanding. In *Proceedings of ICASSP*. IEEE, pp.

[4] R. Gomez *et al.* 2015a. Optimized wavelet-domain filtering under noisy and reverberant conditions. In *Transactions on Signal and Information Processing*. APSIPA, vol. 4.

[5] R. Gomez *et al.* 2015b. Optimizing spectral subtraction and wiener filtering for robust speech recognition in reverberant and noisy conditions. In *Proceedings of ICASSP*. IEEE, pp. 4566–4569.

[6] R. Surikaya *et al.* 2014b. Application of Deep Belief Networks for Natural Language Understanding. In *TASLP*. IEEE, vol. 22 pp. 778–784.

[7] T. Nakamura *et al.* 2011. Multimodal categorization by hierarchical dirichlet process. In *Proceedings of IROS*. IEEE, pp. 1520–1525.

[8] H. Nakajima. 2010. Blind Source Separation with parameter-free adaptive step-size method for robot audition. In *TASLP*. IEEE, vol. 24, pp. 739–761.

[9] R. Schmidt. 1986. Multiple emitter location and signal parameter estimation. In *Trans. Ant. Prop*. IEEE, vol. 34, 276–280.