International Journal of Information, Control and Computer Sciences
ISSN: 2517-9942
Vol:7, No:7, 2013

Automatic Distance Compensation for Robust Voice-based Human-Computer Interaction

Randy Gomez, Keisuke Nakamura, and Kazuhiro Nakadai

Abstract—Distant-talking voice-based HCI system suffers from performance degradation due to mismatch between the acoustic speech (runtime) and the acoustic model (training). Mismatch is caused by the change in the power of the speech signal as observed at the microphones. This change is greatly influenced by the change in distance, affecting speech dynamics inside the room before reaching the microphones. Moreover, as the speech signal is reflected, its acoustical characteristic is also altered by the room properties. In general, power mismatch due to distance is a complex problem. This paper presents a novel approach in dealing with distance-induced mismatch by intelligently sensing instantaneous voice power variation and compensating model parameters. First, the distant-talking speech signal is processed through microphone array processing, and the corresponding distance information is extracted. Distance-sensitive Gaussian Mixture Models (GMMs), pre-trained to capture both speech power and room property are used to predict the optimal distance of the speech source. Consequently, pre-computed statistic priors corresponding to the optimal distance is selected to correct the statistics of the generic model which was frozen during training. Thus, model combinatorics are post-conditioned to match the power of instantaneous speech acoustics at runtime. This results to an improved likelihood in predicting the correct speech command at farther distances. We experiment using real data recorded inside two rooms. Experimental evaluation shows voice recognition performance using our method is more robust to the change in distance compared to the conventional approach. In our experiment, under the most acoustically challenging environment (i.e., Room 2: 2.5 meters), our method achieved 24.2% improvement in recognition performance against the best-performing conventional method.

Keywords—Human Machine Interaction, Human Computer Interaction, Voice Recognition, Acoustic Model Compensation, Acoustic Speech Enhancement.

I. Introduction

Communication is a basic form of human expression. In our day to day lives, we interact with our peers by communicating with them; be it through physical contact or through devices. Nowadays, technology enables us to communicate at our own convenience. With all the different modes to communicate, speech is one of the most natural medium of them all. This reason underscores the importance of harnessing speech in achieving a more effective human-computer interaction (HCI).

The advancement of microprocessor technology has paved the way to a more efficient and fast computers. Now is the time of computing power sufficient enough to bring into reality applications that once were just science fiction. Consequently, device sensors have flooded the market which fuels the demand for development of HCI applications. This trend is gaining traction and is expected to maintain its momentum in the next decade [1]. In recent years, we have been witnessing the advent of immersive gaming features of consoles equipped with state-of-the-art HCI. The once hand-held controller-restricted gaming has evolved to controller-less experience using gestures and bodily movements for gaming interaction. We see smartphones equipped with several sensors (e.g. gyroscope, GPS, proximity, etc.) that interpret user's intention, enabling the hand-held device towards smart interaction. Another prominent household appliance that benefitted from the makeover is the monitor display for television (TV) set. It is currently referred to as smart TV or smart display, to emphasize a human-centric design capable of interacting with users through smartphones, gestures, etc. smart TV sports a stark contrast to the conventional TV, in which the interaction experience was limited to the use of bulky remote controls. In the near future, most of the general appliances in the household will be equipped with HCI features.

Recently, the feat in HCI experience has further pushed the envelope with the inclusion of the speech modality. After all, speech is human's preferred mode of communicating. This provision allows users to issue speech-based command for interaction using their hand-held devices. Apparently, HCI experience is very much dependent on the voice recognition performance of the system. Surveys on the use of speech-based technology in HCI show a correlation between voice recognition performance (e.g. speech command are correctly recognized) and the HCI experience [2]. In particular, the level of satisfaction is high when the voice commands are recognized correctly, and users' dissatisfactions are imminent when voice recognition fails in which users have to repeat the voice command for a number of times [2]. Overall, the addition of voice recognition feature has gained wide acceptance among users.

When using hand-held devices such as smartphones, the quality of the speech signal is in good acoustic condition when observed by the microphone, as the device is held closely to the mouth. In the case of smart TV and smart displays, we need a hands-free system since the user is of considerable distance away from the device [3][4]. In this scenario, the speech signal is susceptible to acoustic changes as it travels in free-space, and reflected within the room enclosure. This phenomenon drastically affects recognition performance.

Model-based voice recognition systems employ acoustic models trained with acoustic speech from speech database [5][6]. Then, the pre-trained acoustic model is used against the speech command at runtime. Such system is very sensitive to mismatch, especially when speech acoustics at runtime condition is different from the condition when the model was trained [7]. Degradation in voice recognition performance

Honda Research Institute Japan (HRI-JP) Co. Ltd., Wako-shi, Saitama, 351-0188 JAPAN.

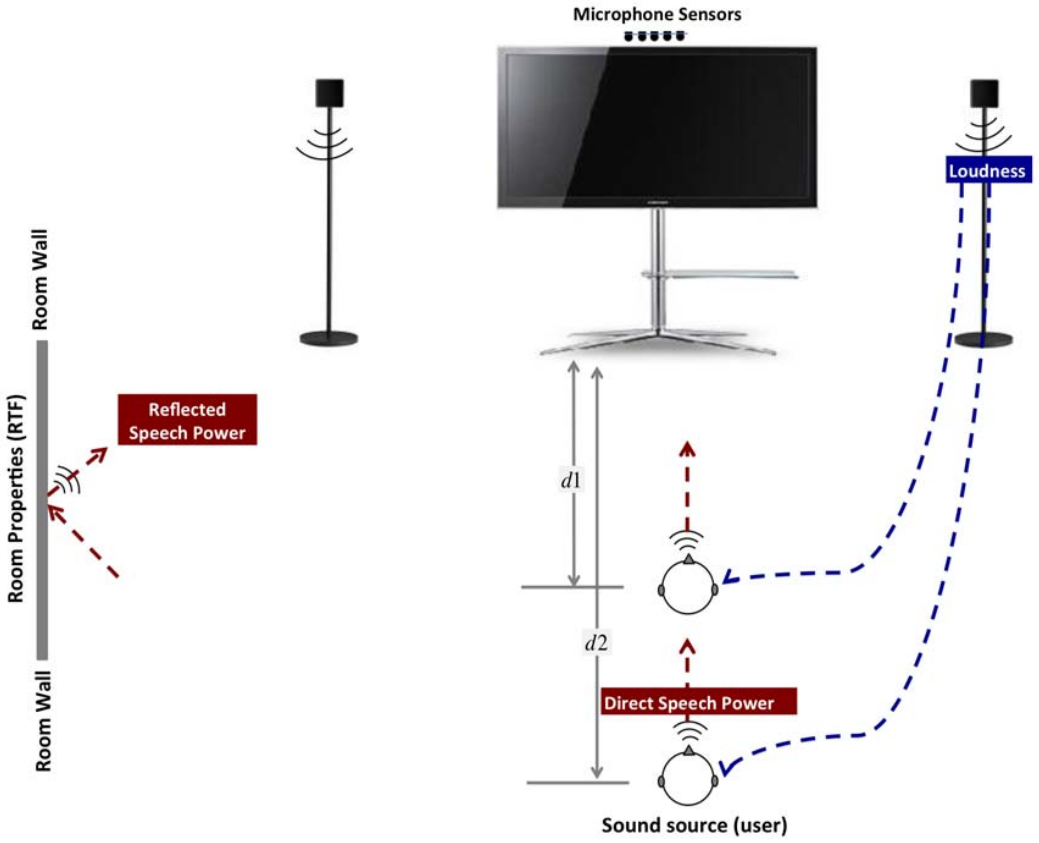

Fig. 1. Human computer interaction (HCI) set-up in a distant-talking and enclosed environment

is imminent at runtime depending on the severity of the mismatch. There exist many kinds of mismatch such as noise, speaking styles and speaker variability, among others [8]-[14]. Since research works on these particular issues have been intensive, in this work, we focused on the mismatch caused by the variation of the observed acoustic speech power due to distance. Specifically, in this paper, we address the drop in speech power as the user moves away from the system inside an enclosed environment (e.g. room). In particular, we addressed the effect of speech power vis-a-vis the room properties.

The organization of the paper is as follows; in Sec. II we introduce the problem of acoustic mismatch inside an enclosed environment, and the conventional voice-based HCI system in Sec. III. In Sec. IV, we present the proposed method that addresses both speech and model mismatch, followed by the experimental set-up in Sec. V. Recognition performance results in real HCI environment is presented in Sec. VI. Lastly, we conclude this paper in Sec. VII.

II. Acoustics Mismatch in Human Computer Interaction Inside an Enclosed Environment

Consider the HCI set-up in Fig. 1. Interaction between the smart display and the user is initiated with a speech command. As a result, the system replies via synthetic speech through the speakers which are hooked to an amplifier with variable loudness control. It is assumed that the user is of considerable distance, and the user changes position d, relative to the smart display. In an enclosed room as shown in Fig. 1, both the direct speech and the reflected speech are observed at the microphone sensors [15]-[18]. As the speech is reflected from the walls and ceilings, its acoustical characteristic is altered by the room properties which is characterized by the room transfer function (RTF) [19]. Hence, the observed speech power at the microphone sensors is difficult to model as it includes both the direct and the reflected speech. However, these two are dependent on the distance d, between the smart display and the source, so we can associate the observed power at the microphone sensors to the distance d for simplicity. Thus, d implicitly describes the observed power at the microphones. Furthermore, as d increases, the observed power also decreases, thus, voice recognition performance is inferior at $d2$ than in $d1$ in Fig. 1. Conventional methods [20]-[26] fall short to mitigate the degradation of voice recognition performance as a result of distance variation due to the following :

- *When employing microphone array technology, conventional methods deal primarily on the effects of the room properties and not on the change in power.*

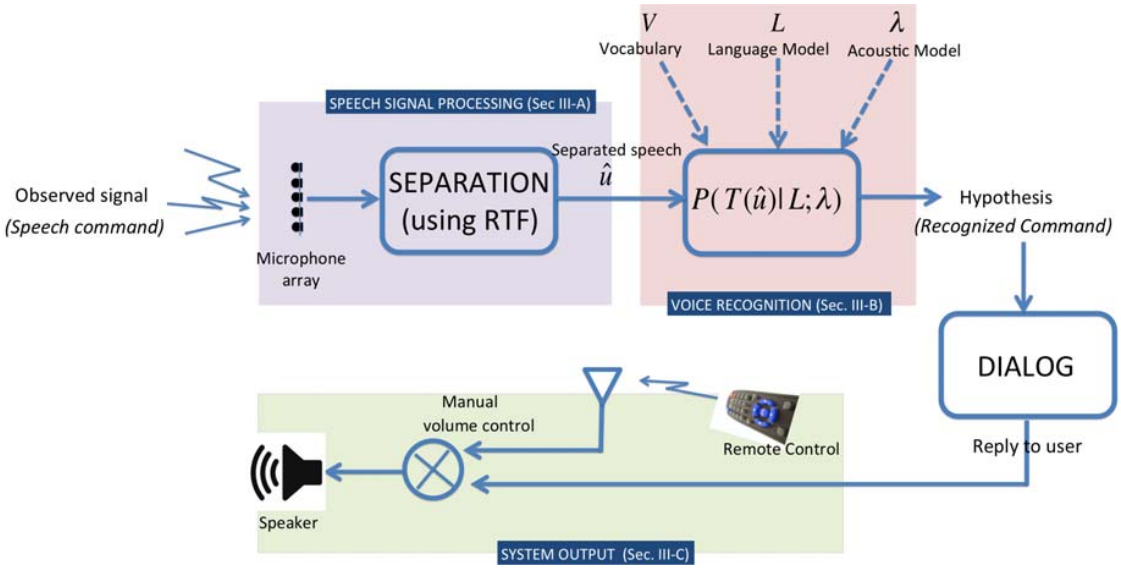

Fig. 2. Conventional method that addresses only the acoustic speech while leaving the model out of the design problem.

- *Limited focus to a fixed-distance design approach between the speech source and the system. This generic design is assumed to work with different distances resulting to mediocre performance.*

- *Pre-trained generic model is assumed to work at runtime with different d. Thus, mismatch between the model and the observed speech is ignored in the conventional design. This assumption fails when both training and runtime conditions are completely dissimilar especially when d varies considerably like that in Fig. 1.*

This paper addresses the effect of distance-induced mismatch in a holistic approach, by considering the synergy of the mismatch between the observed speech and the pre-trained acoustic model. Unlike the conventional methods [20]-[26], we also compensate power issues in the model level at runtime, depending on the acoustics of the observed speech. Thus, the proposed method is robust to the degradation of voice recognition performance as a function of d. Moreover, the system can automatically adjust the loudness of the speaker depending on the distance of the user. We evaluate the effectiveness of the proposed method using real data and show it outperforms the conventional methods in terms or recognition performance, which is a main indicator of a users' satisfaction in speech-based HCI system [2].

III. Conventional HCI System

The block diagram of the conventional HCI system is shown in Fig. 2. The incident speech as observed by the microphones is processed using speech signal processing technique. The separated speech $\hat{u}$ is used as input to the voice recognition system, and the recognized command is fed into a dialog manager that generates the audio reply of the system to the speaker. Volume control of the speaker is manually set by the user.

A. Speech Signal Processing

1) Microphone Array Processing: The use of multiple microphone sensors provides a more enhanced separated speech signal for improved signal-to-noise ratio prior to voice recognition. Suppose that there are N sources and M ($\geq N$) microphones. Let $\boldsymbol{u}(\omega)$ denote the input acoustic signal of N sources in frequency domain, described as $\boldsymbol{u}(\omega) = [u_1(\omega), \cdots, u_N(\omega)]^T$, where T represents a transpose operator. $\boldsymbol{x}(\omega) = [x_1(\omega), \cdots, x_M(\omega)]^T$ are the signals received by M microphones. The model for microphone array signal processing is described as follows:

$$\boldsymbol{x}(\omega) = \boldsymbol{A}(\omega)\boldsymbol{u}(\omega) + \boldsymbol{n}(\omega), \tag{1}$$

where $\boldsymbol{A}(\omega) \in \mathbb{C}^{M \times N}$ is a *Room Transfer Function (RTF)* matrix between a microphone array and sound sources; $\boldsymbol{n}(\omega)$ is an additive noise, which is assumed to be statistically independent of $\boldsymbol{u}(\omega)$. The RTF contains information regarding the characteristics and properties of the room, and this information is used to reflect room characteristics in the separated signal discussed below.

2) Separation via Room Transfer Function: The sound sources are spatially separated by a hybrid algorithm of beamforming and blind separation called *Geometrically constrained High-order Decorrelation based Source Separation (GHDSS)*. Using the input signal $\boldsymbol{x}(\omega)$, $\hat{\boldsymbol{u}}(\omega)$ is usually defined by $\hat{\boldsymbol{u}}(\omega) = \boldsymbol{V}(\omega)\boldsymbol{x}(\omega)$ in frequency domain, where $\boldsymbol{V}(\omega)$ is called a separation matrix. GHDSS updates $\boldsymbol{V}(\omega)$ so that it can correctly estimate $\boldsymbol{u}(\omega)$ in Eq. (1) by $\hat{\boldsymbol{u}}(\omega)$. In order to estimate $\boldsymbol{V}(\omega)$, GHDSS introduces two cost functions, that is, separation sharpness (J_{SS}) and geometric constraints (J_{GC}):

$$J_{SS}(\boldsymbol{V}(\omega)) = \|\phi(\hat{\boldsymbol{u}}(\omega))\hat{\boldsymbol{u}}^H(\omega) - \mathrm{diag}[\phi(\hat{\boldsymbol{u}}(\omega))\hat{\boldsymbol{u}}^H(\omega)]\|^2$$
$$J_{GC}(\boldsymbol{V}(\omega)) = \|\mathrm{diag}[\boldsymbol{V}(\omega)\boldsymbol{A}(\omega) - \boldsymbol{I}]\|^2$$

where $\|\cdot\|^2$ indicates the Frobenius norm, $\mathrm{diag}[\cdot]$ is the diagonal operator, and H represents the conjugate transpose

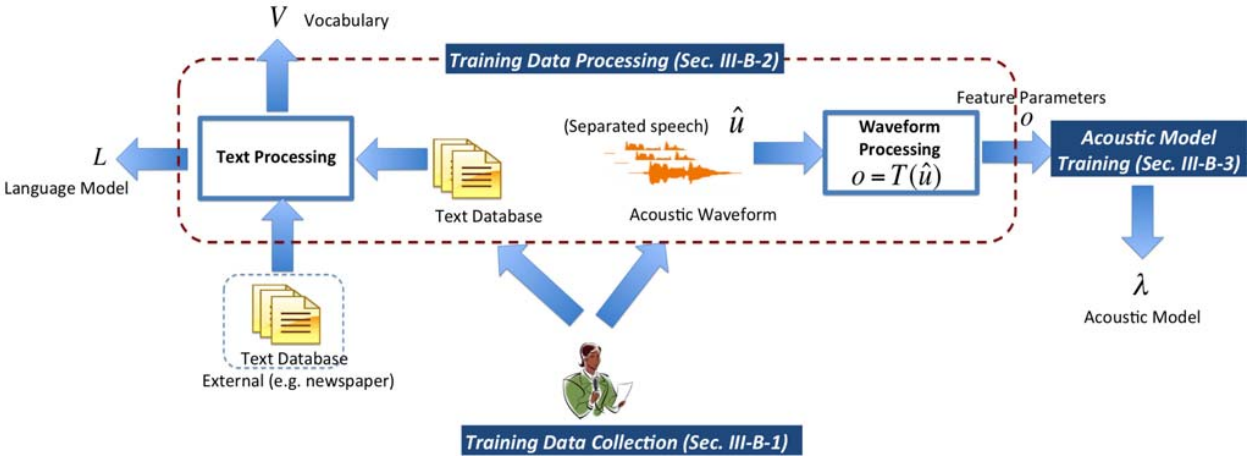

Fig. 3. Block diagram of the processes needed for voice recognition (prepared offline).

operator. For a nonlinear function, $\phi(\hat{\boldsymbol{u}}(\omega))$, we selected a hyperbolic-tangent-based function [27] in this paper. Since the best $\boldsymbol{V}(\omega)$ is always changing in the real world, $\boldsymbol{V}(\omega)$ is adaptively updated as described in [28]. Consequently, the separated signal $\hat{u}(\omega)$ is extracted [28].

B. Voice Recognition

In recognizing a voice command, the system converts the acoustic speech (separated speech) into text format referred to as hypothesis. The three important components in voice recognition prepared offline are,

- Vocabulary *The list of words that are defined in a corresponding task.*
- *Acoustic Model The model that capture the statistical characteristics of the sounds in the vocabulary. In our system we use the hidden markov model (HMM)[5][29].*
- *Language Model The model that contains the statistics of word sequences in a given task. Aids the acoustic model to generate the most likely hypothesis.*

The process of preparing these components is shown in Fig. 3.

1) Training Data Collection: The speech waveform and its corresponding text transcriptions are required prior to training. Speech utterances are collected from different speakers to ensure variety of speaking styles for a wide coverage of possible acoustical variations. Moreover, It is important that all possible sound units are well represented (i.e., phonetically balanced) in the training database to ensure a sufficient amount of data prior to acoustic model training. The recorded speech data is required to be in digital format. When recording the speech data, it is important to consider recording parameters such as sampling frequency, resolution (bits), and the type of microphone being used, as these may cause mismatch and affect performance.

2) Training Data Processing:

- Text Processing:
  The accompanying text is used to associate the acoustical sound (one-to-one correspondence). Using the generated transcripts during the recording phase, we extract the words or vocabularies that come along with the speech waveform. Given both the transcripts and the wave data, supervised acoustic model training can be performed. Voice recognition follows the same principle concerning text queries, except for using speech as input. The more text data used in language modelling, the better chance of answering the queries. Thus, it is beneficial to gather text data from external sources to model the most probable queries. Language model contains the statistics of word sequences. Specifically, it puts probability measure over strings found in a document which generates the most likely query. In short, language model aids the acoustic model to generate the query when using speech input.

- Waveform Processing:
  The speech waveform contains a huge amount of information in the time domain, and most of these do not bear meaningful information. Thus, we need to process these data and extract observation vectors through spectral analysis [30]. By performing spectral analysis, the time-domain speech is represented in a more compact and meaningful way. First, speech signal is emphasized to flatten its response spectrally [31]. This makes sure that all frequency components of the signal are treated equally. Then, the time domain-signal is blocked into frames. Each frame corresponds to a meaningful representation of sound (e.g. a phone) with a defined duration [32]. The adverse effects of blocking the data into frames is minimized through windowing. Finally cepstral analysis is performed for each frame to capture speech features that best describes the speech characteristics [30]-[31]. These low-dimensional features (as compared to the time-domain counterpart) are the observation vectors used

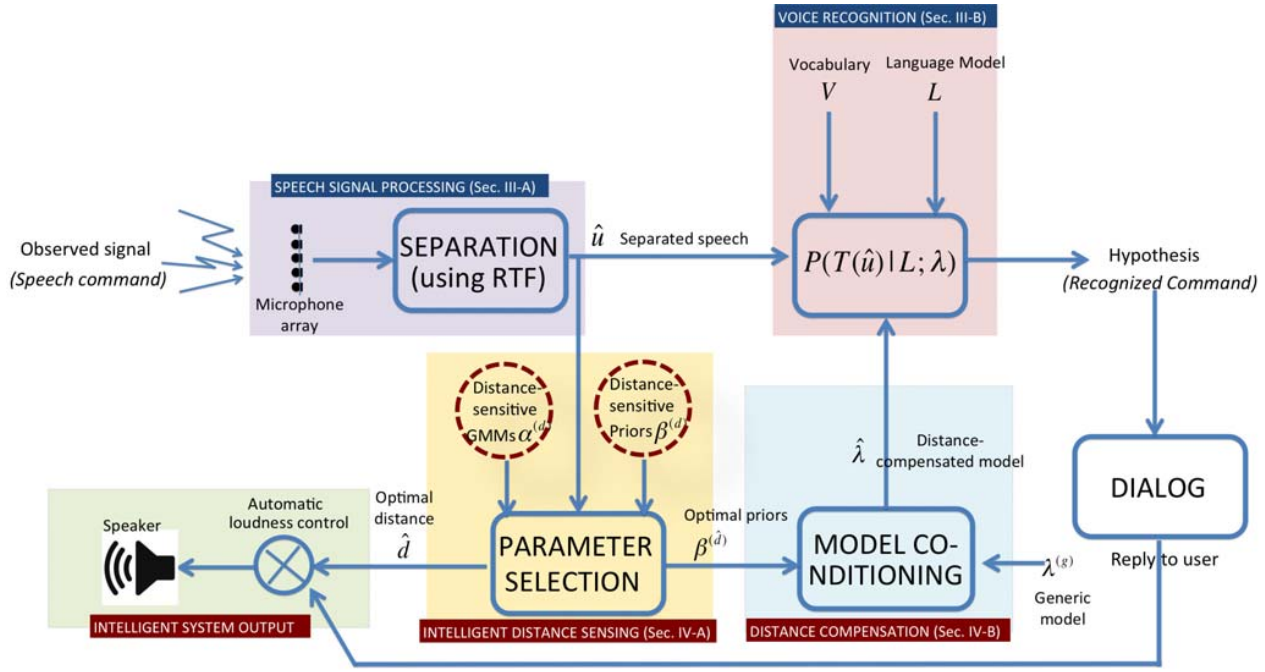

Fig. 4. Proposed method reflecting acoustical change in the separated speech to the acoustic model by means of compensation using distance-sensitive priors.

to train the acoustic model. The transformation of the separated speech to features can be viewed as the transformation $o = T(\hat{u})$.

3) Acoustic Model Training: The acoustic model is the heart of a model-based system [33]. Specifically, in our method we use the Hidden Markov Model (HMM). The HMM, contains the statistics of the speech data. Training an acoustic model means optimizing the the parameters of the HMM

$$\hat{\lambda} = \arg\max_{\lambda} \prod_{r=1}^{R} P(\mathbf{o}_r|\mathbf{V};\lambda), \tag{2}$$

where λ is unknown model parameters and $\mathbf{V}$ is the word sequence. $\mathbf{o}_r$ is the r-th training observation vector derived from the speech utterance. The optimization of λ has to be carried out separately for each of the training utterance and the process is iterated until a reasonable performance is achieved. Thus, iterative model training is employed. There are several algorithms to train an acoustic model. In our case, we use the Expectation Maximization approach through Baum Welch [5][33]. Specifically, we optimized the the following parameters of λ:

$$C_{im} = \frac{L_{im}}{\sum_{m=1}^{M} L_{im}}, \tag{3}$$

$$\mu_{im} = \frac{\boldsymbol{m}_{im}}{L_{im}}, \tag{4}$$

$$\Sigma_{im} = \frac{\boldsymbol{v}_{im}}{L_{im}} - \mu_{im}\mu_{im}^{T}, \tag{5}$$

$$a_{ij} = \frac{L_{ij}}{\sum_{j=1}^{J} L_{ij}}, \tag{6}$$

where C_{im}, μ_{im}, Σ_{im}, and a_{ij} are the mixture weight, mean, covariance matrix and updated transition probability respectively. m denotes the mixture while i and j signify the state (i is the current state). L_{im}, L_{ij}, $\boldsymbol{m}_{im}$, $\boldsymbol{v}_{im}$ are the accumulated mixture occupancy, state transition occupancy, mean statistics and variance statistics, respectively. In this paper we refer to L_{im}, L_{ij}, $\boldsymbol{m}_{im}$, $\boldsymbol{v}_{im}$ as priors, and collectively denoted as β which will be used for model conditioning (Sec. IV-B). The details of these statistics are found in [5].

C. System Output

As soon as the speech command is recognized, the hypothesis is then processed by the dialog manager [34]-[35] that interprets the command to a corresponding action by the system. In our case, an audio reply from the system in the form of synthetic speech, is being fed to an amplifier and then to the speaker. In this set up, the system does not automatically adjust the volume of the amplifier as a function of the distance from the user. Thus, the user has to manually set speaker loudness level for each different position.

The problem with the conventional approach in Fig. 2 is that it only addresses the effect of the RTF (room properties) [26] in the separated speech, and does not take into consideration the acoustic model λ. This is one major cause of mismatch as the former may not be of the same condition as the pre-trained model (i.e., different room properties and power level variation).

IV. Proposed Method

The proposed method is shown in Fig. 4. After signal processing, distance information is extracted from the separated

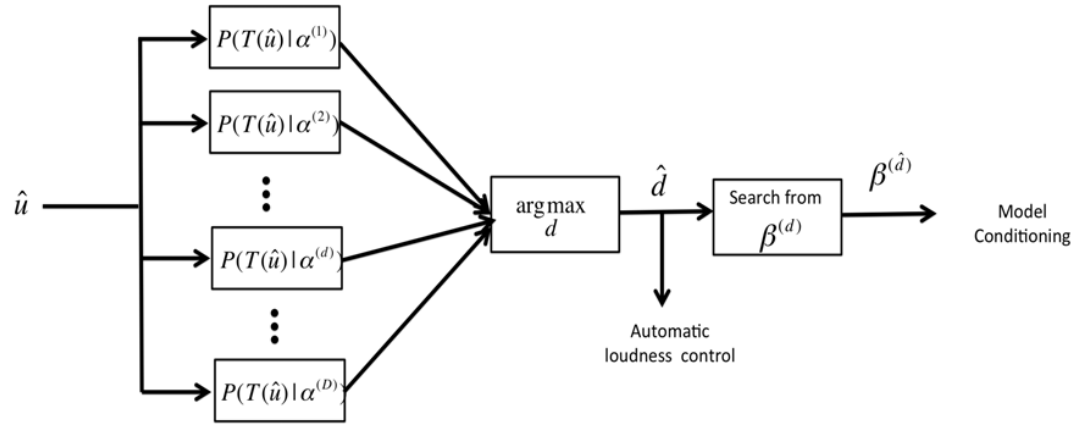

Fig. 5. Parameter selection using distance-sensitive gaussian mixture model classifiers.

speech signal at runtime, which is used for parameter selection needed to condition the generic model. Consequently, the system automatically adjusts the loudness level as it interacts through the speaker.

A. Intelligent Distance Sensing

The distance d is associated to the room acoustics in Sec. I. And it implicitly describes the observed power in the microphone array. Thus, by identifying d we can predict the effects of power mismatch from a particular source location as observed by the microphone. We note that the RTF affects the observed power as it influences the acoustics of the reflected speech. RTF $A(\omega)$ can be acquired through physical measurement as described in [36], and this has been validated to work in [20] [21]. However, it is impractical to measure room RTF for every room where the system is deployed; and when dealing with bigger rooms, the RTF may vary as a function of distance. Conveniently, there exist several techniques in approximating the RTF $\hat{A}(\omega)$ by modeling the effects of the reflection using an exponentially decaying function as introduced in [19][37]. Furthermore, we have also shown in [38] the technique of compensating the effect of speech power as a function of distance d with an interpolator $f(d)$. Thus, we can synthetically recreate a speech signal that contains both RTF (room property) with different speech power as a function of distance d by

$$\hat{u}^{(d)} = s(\omega)f(d)\hat{A}(\omega,d), \tag{7}$$

where $s(\omega)$ is a clean speech signal using closed-talking microphone, $f(d)$ is the power compensation technique we developed in [38] and $\hat{A}(\omega)$ is the RTF estimate [19][37]. Eq. (7) models the separated speech which is analogous to the actual separated speech when processed by the microphone array. The only difference here is that we can simulate the separated speech using Eq. (7) without physically recording it. Specifically, we use Eq. (7) to synthetically generate speech data at different d which will be used to train or distance-sensitive classifiers and priors.

- **Distance-sensitive Gaussian Mixture Models (GMM)**
  To effectively identify d, we design a GMM-based d classifier $\alpha^{(d)}$. Prior to the actual classification, speech data $\hat{u}^{(d)}$ for $d = 1...D$ are synthetically generated using Eq. (7) and used to train GMMs $\alpha^{(d)}$. We use a large-mixture GMM (i.e. 256 mix.) to better capture the room characteristics and speech power variation at distance d. The choice of the number of gaussian mixtures is explained using Table I in Sec. V. In this work, we experimentally set the step size of d to 0.5 m. covering from 0.5m to 2.5m.
  Although cameras or other device such as kinect are good candidates to identify the distance between the smart display and the user; these are sensitive to lighting conditions, and may not work especially in the dark. Moreover, distance d in this paper is not solely limited to the distance measurement per se, but it is also associated to room information (i.e. property) which is embedded in the speech signal caused by reflections to the walls and ceilings. And room information (property) cannot be obtained through the use of camera. Thus, we focus only with speech modality in this paper.

- **Distance-sensitive Priors**
  Using the same synthetic data $\hat{u}^{(d)}$ from Eq. (7) we calculate the priors $\beta^{(d)}$ which consists of $L_{im}^{(d)}$, $L_{ij}^{(d)}$, $\boldsymbol{m}_{im}^{(d)}$, $\boldsymbol{v}_{I'm}^{(d)}$ and stored in the priors database for use during model compensation at runtime. These are the accumulated mixture occupancy, state transition occupancy, mean statistics and variance statistics, respectively. Thus, a total of $d = D$ classes of priors. We note that both the effects of power and RTF at a particular d are infused in these statistics. Unlike the GMMs, these statistics are prepared for HMMs since these will be used in conditioning the HMM for voice recognition.

- **Parameter Selection**
  Optimal Parameter selection based on likelihood score is

TABLE I
RESULTS IN IDENTIFYING THE CORRECT DISTANCE d USING DISTANCE-SENSITIVE GAUSSIAN MIXTURE MODEL (GMM) CLASSIFICATION.

	Room 1			Room 2		
No. of Gaussian Mixtures	Loc 1	Loc 2	Loc 3	Loc 1	Loc 2	Loc 3
2 mixtures	10 %	8 %	9 %	3 %	1 %	2 %
4 mixtures	18 %	15 %	19 %	5 %	4 %	5 %
8 mixtures	29 %	27 %	27 %	12 %	15 %	14 %
16 mixtures	40 %	42 %	40 %	24 %	26 %	25 %
32 mixtures	57 %	55 %	59 %	37 %	40 %	38 %
64 mixtures	79 %	80 %	76 %	52 %	57 %	55 %
128 mixtures	90 %	91 %	88 %	68 %	71 %	70 %
256 mixtures	**98 %**	**95 %**	**97 %**	**80 %**	**83 %**	**81 %**
512 mixtures	98 %	96 %	97 %	81 %	84 %	81 %

employed using the pre-trained GMMs, and the classification process is shown in Fig. 5. The separated speech $\hat{u}$ from an unknown sound source location at runtime is transformed to feature vectors $T(\hat{u})$, and then fully evaluated against the GMMs $\alpha^{(d)}$ for $d = 1..D$. Subsequently, the argument $\hat{d}$ that maximizes the likelihood score is selected and used to adjust the volume control for the system's speaker output. Consequently, $\hat{d}$ is used to select the corresponding prior $\beta^{(\hat{d})}$ from the priors database which will be used in conditioning the generic HMM $\lambda^{(g)}$ prior to voice recognition.

B. Distance Compensation

- **Generic HMM**
  A generic model is trained as described by the training procedure in Eq. (2) using a clean speech recorded from close-talking microphone. We note that close-talking condition is usually the standard practice in most speech databases available (as opposed to distant-talking). The resulting generic HMM $\lambda^{(g)}$ is void of RTF and power variation information. We use this as our base model.

- **Model Conditioning**
  Conditioning the HMM $\lambda^{(g)}$ means reflecting some statistics from $\beta^{(\hat{d})}$ to $\lambda^{(g)}$. In our case, $\beta^{(\hat{d})}$ is employed in a form of bias to the model parameters of $\lambda^{(g)}$. Thus, Eq. (3)-(6) are modified and the parameters of $\hat{\lambda}$ (distance-compensated model) become

$$\hat{C}_{im} = \frac{L_{im}^{(g)} + \tau L_{im}^{(\hat{d})}}{\sum_{m=1}^{M} L_{im}^{(g)} + \tau L_{im}^{(\hat{d})}}, \tag{8}$$

$$\hat{\mu}_{im} = \frac{\boldsymbol{m}_{im}^{(g)} + \tau \boldsymbol{m}_{im}^{(\hat{d})}}{L_{im}^{(g)} + \tau L_{im}^{(\hat{d})}}, \tag{9}$$

$$\hat{\Sigma}_{im} = \frac{\boldsymbol{v}_{im}^{(g)} + \tau \boldsymbol{v}_{im}^{(\hat{d})}}{L_{im}^{(g)} + \tau L_{im}^{(\hat{d})}} - \mu_{im}^{(d)} \mu_{im}^{(d)^T}, \tag{10}$$

$$\hat{a}_{ij} = \frac{L_{ij}^{(g)} + \tau L_{ij}^{(\hat{d})}}{\sum_{j=1}^{J} L_{ij}^{(g)} + \tau L_{ij}^{(\hat{d})}}, \tag{11}$$

where $\hat{C}_{im}$, $\hat{\mu}_{im}$, $\hat{\Sigma}_{im}$, and $\hat{a}_{ij}$ are the distance-compensated mixture weight, mean, covariance matrix and updated transition probability respectively. m denotes the mixture while i and j signify the state (i is the current state). $L_{im}^{(\hat{d})}$, $L_{ij}^{(\hat{d})}$, $\boldsymbol{m}_{im}^{(\hat{d})}$, $\boldsymbol{v}_{im}^{(\hat{d})}$ are the priors $\beta^{(\hat{d})}$ which were pre-computed in advance (to be used as bias); τ is the multiplying constant that was experimentally determined. In our experiment we used $\tau = 0.1$.

The bias has two significant contributions. First, it has the effect of shifting the statistics of the generic model $\lambda^{(g)}$ which was frozen during training, closer to the runtime condition as depicted by the distance $\hat{d}$ (Note that we associate overall room conditions with d). The combinatorics of the model $\lambda^{(g)}$ was left open after training and dynamically resolved at runtime depending on the the current acoustics condition. In doing so, the mismatch between training and runtime condition is minimized in the HMM level. Second, it is important to underscore that despite the perturbation caused by the change in distance d, the HMM should not lose its discriminative property to effectively recognize speech. The base parameter $\lambda^{(g)}$ is considered to be "complete" model, being trained from reliable data as far as speech recognition is concerned. In the bias mechanism Eq. (8)-(11), we can maintain this status quo by setting $\tau << 1$ rendering $\lambda^{(g)}$ to be the dominant statistics. Thus, we can infuse environment information through $\beta^{(\hat{d})}$ while maintaining the discriminatory property to recognize speech in the new model (i.e., distance-compensated model).

Although it is possible to correct the model through the use of pre-existing adaptation techniques. Users are asked to enrol adaptation data (apart from training database) which is used to adapt the acoustic model. This approach is impractical since there are many possible values of d requiring separate sets of adaptation data. Moreover, it takes some processing time to perform model adaptation. Our method does not require any adaptation data from the user (we generate synthetic data from training database) and it is executed at runtime. Thus, our method is more practical and convenient.

V. EXPERIMENTAL SET-UP

A. Training and Testing Database

The close-talking clean speech training database is from the Japanese Newspaper Article Sentence (JNAS) corpus. This is

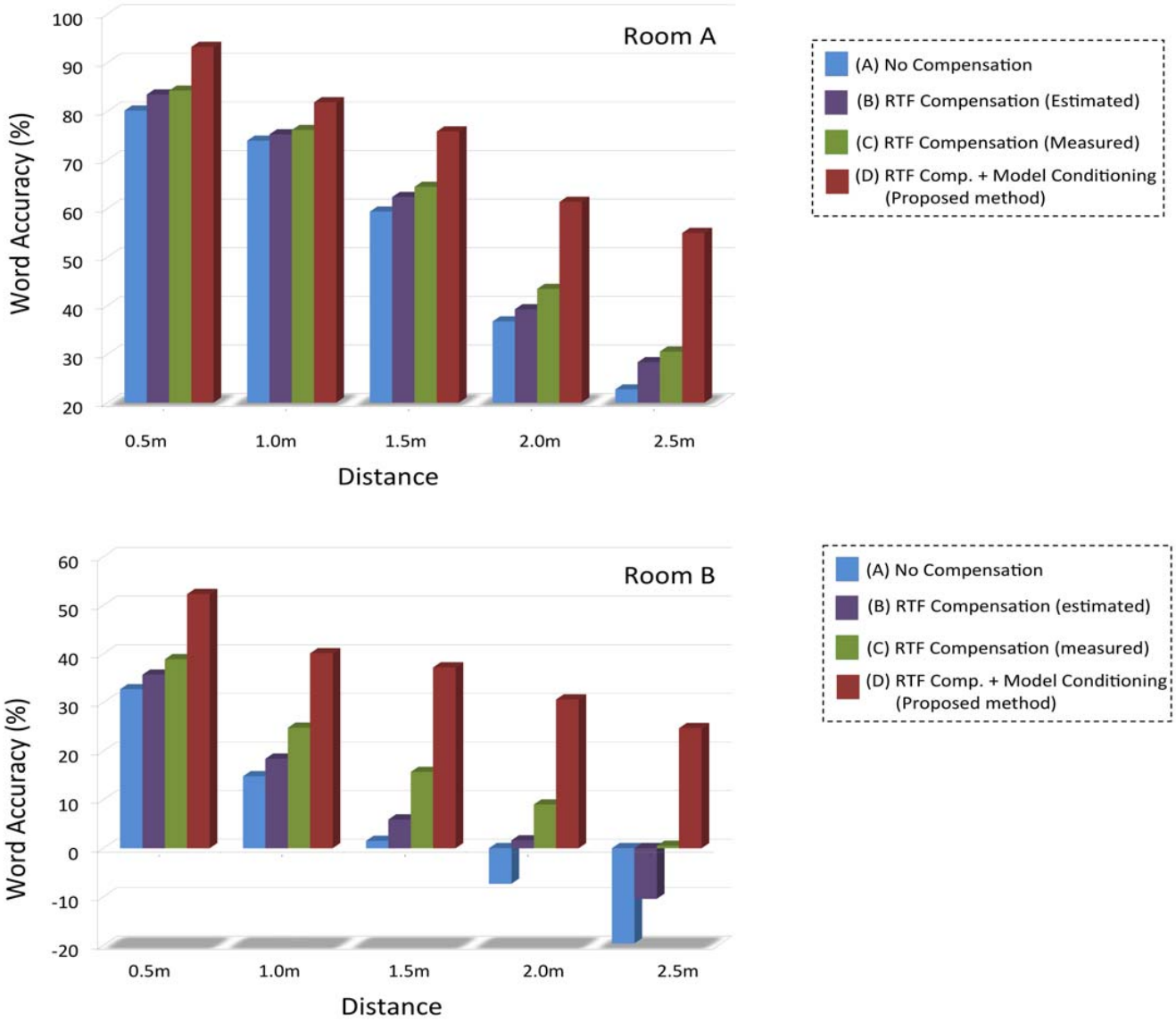

Fig. 6. Basic Recognition Results.

used to train the distance-sensitive GMMs; compute the priors; and train the generic HMMs. Recognition experiments are carried out on our HCI task with 1K-word vocabulary. The open test set is composed of 50 utterances coming from 10 speakers and the language model is a standard word trigram model. The acoustic model is a phonetically tied mixture (PTM) HMMs with 8256 Gaussians in total. Test experiment is conducted using actual HCI set-up shown in Fig 1. The microphone array is embedded on top of the smart display.

The test set is recorded inside two different rooms (Room 1 and Room 2) with reverberation time of 240 ms (mild echo) and 640 ms (strong echo), respectively. Thus, Room 2 is worse than Room 1 in terms of acoustic condition. Inside the room, there is environment noise (i.e., computer noise) and the sound that comes from the speakers of the smart display. The microphone array is positioned to minimize the impact of the environmental noise. A total of five different radial distances d are considered $\{0.5\text{m}, 1.0\text{m}, 1.5\text{m}, 2.0\text{m}, 2.5\text{m}\}$. Thus, the test set is recorded separately for different distances d in Room 1 and Room 2.

B. GMM Classification Performance

Distance identification is vital in selecting the appropriate priors for model compensation and to set correct loudness level. We show in Table 1 the classification rate (correctly identifying d) with different gaussian mixtures and in different rooms (Room 1 and Room 2). Moreover, in each room, we randomly changed the location of the smart display 3 times to check the consistency of the results at different locations

(i.e., Loc. 1, Loc. 2 and Loc. 3). We have dyadically increased the gaussian mixture in each training from 2 mixtures up to 512 mixtures. With lesser mixtures, the classifier is unable to discriminate the acoustical dynamics for every change in the distance d. This results to poor d identification rate. As the mixtures are increased, the identification rate improves and saturates at 256 mixtures. This means that with increased mixtures, the classifier is more capable of capturing the room dynamics as a function of d. Room 1 has better identification rate than Room 2 because it is less echoic than the latter.

VI. Results and Discussions

Recognition performance is one of the most important measures for an effective HCI [2]. We show in Fig. 6 the performance of the different methods used in our experiment for both Room A and Room B, with d=$\{0.5\text{m}, 1.0\text{m}, 1.5\text{m}, 2.0\text{m}, 2.5\text{m}\}$. (A) is the performance when there is no compensation in effect. (B) and (C) are the results when the acoustics speech is processed with RTF information during the microphone array processing (Conventional method) [26]. Specifically, (B) employs an estimated RTF using a mathematical model [19][37], while (C) uses a physical measurement of the RTF [36]. Although (B) and (C) improved the recognition performance as compared to (A), the improvement is minimal. This is attributed to the fact that compensating the acoustic speech alone does not minimize the mismatch with the acoustic model. Lastly, the proposed method is shown in (D), where a consistent improvement in recognition performance is achieved in both rooms. Again, Room 1 has better performance than Room 2 because it is less echoic than the latter.

The proposed method is more robust to changes in distance d as opposed to the conventional approach in (B) and (C) as shown by the significant improvement of the word accuracy. This result is a manifestation that addressing the acoustic model in accordance to actual acoustic condition of the speech, effectively minimizes the mismatch between the two. Thus, improving the likelihood of recognizing the speech command at various distances d. This makes sense because during training, the model combinatorics are frozen to "training condition" and there is no assurance that this condition is the same during actual use at runtime. Since voice recognition requires both the speech data and the acoustic model altogether, it is important to consider mismatch issue between these two jointly. Thus, the synergetic effect of compensation through RTF (acoustic speech) and through the model brings significant impact to the improvement of the recognition performance.

The robustness in voice recognition performance as a function of d results to an increased in HCI experience satisfaction among users. Moreover, the users preferred the automatic loudness control of the proposed method as opposed to the manual setting of the conventional method. All the participants respond positively when asked regarding the automatic loudness control of the system.

VII. Conclusion

We have presented an approach that addresses speech power variation in a distant-talking environment. By associating the acoustical dynamics of the observed speech to the distance d, we have simplified the analysis of tackling the issue of power variation. Modelling the speech power as observed by the microphone array is a very difficult task due to the room acoustics. And we simplified this procedure by identifying the distance d instead. In this paper, we have shown the method of synthetically generating the training data from a mathematical model that best describes the acoustic speech (i.e., both the effects of the RTF and power). From these data, we are able to identify the acoustical condition of the actual speech utterance at runtime. Moreover, we used the same synthetic data in computing the priors, used to compensate the acoustic model dynamically at runtime.

We have significantly reduced the mismatch between training (model) and runtime acoustic condition that renders the recognition performance of the system robust to the change in distance d as opposed to the conventional methods. We note that in real HCI environment conditions d often varies, and room acoustic condition is unpredictable. In the future, we will focus on further improving performance from very far distances (i.e., 1.5m - 2.5m). Moreover, investigation in further improving performance in an acoustically challenged environment (i.e., Room 2) will be conducted. We note that issues may not be solely due to the change in distance, thus, further study is needed.

References

[1] "http://www.gartner.com" *Information technology research and advisory company*

[2] R. Gomez, T. Kawahara, K. Nakamura and K. Nakadai "Multi-party Human-Robot Interaction with Distant-Talking Speech Recognition" *In Proceedings IEEE Human Robot Interaction*, 2012

[3] M. Seltzer, "Speech-Recognizer-Based Optimization for Microphone Array Processing" *IEEE Signal Processing Letters*, Vol. 10, No. 3, 2003

[4] M. Seltzer and R. Stern, "Subband Likelihood-Maximizing Beamforming for Speech Recognition in Reverberant Environments" *IEEE Trans. on Audio, Speech, and Lang. Proc.*, Vol. 14, No. 6, 2006

[5] The HTK documentation http://htk.eng.cam.ac.uk/docs/docs.shtml

[6] Kaifu Lee "Automatic Speech Recogntion – The Development of SPHINX System" *Kluwer Academic Publishers*, Boston, 1989

[7] R. Gomez, J. Even, H. Saruwatari, and K. Shikano, "Rapid Unsupervised Speaker Adaptation Robust in Reverberant Environment Conditions" *In Proceedings Interspeech*, 2008

[8] L. Lee and R. Rose, "Speaker Normalization using Efficient Frequency Warping Procedures" *In Proceedings IEEE Int. Conf. Acoust., Speech, Signal Proc. ICASSP*, pp 353-356, 1996

[9] D.Pye and P.C.Woodland "Experiments in Speaker Normalisation and Adaptation for Large Vocabulary Speech Recognition" *In Proceedings IEEE Int. Conf. Acoust., Speech, Signal Proc. ICASSP*, pp 1047-1050, 1997

[10] A. Baba, S. Yoshizawa, A. Lee, H. Saruwatari, and K. Shikano, "Elderly Acoustic Model fro Large Vocabulary Continuous Speech Recogntion" *In Proceedings EUROSPEECH*, 2001

[11] C. Huang, T. Chen, S. Li and JL. Zhou "Analysis of Speaker Variability" *In Proceedings EUROSPEECH*, 2001

[12] D. Pye and P.C. Woodland "Experiments in Speaker Normalisation and Adaptation for Large Vocabulary Adaptation" *In Proceedings IEEE Int. Conf. Acoust., Speech, Signal Proc. ICASSP*, 1997

[13] Guiliani and Gerosa "Invetsigating Recognition of Children's Speech " *In Proceedings IEEE Int. Conf. Acoust., Speech, Signal Proc. ICASSP*, 2003

[14] R. Gomez and T. Kawahara "Denoising Using Optimized Wavelet Filtering for Automatic Speech Recognition" *In Proceedings Interspeech*, 2011

[15] K. Kinoshita , T. Nakatani and M. Miyoshi, "Efficient Blind Dereverberation Framework for Automatic Speech Recognition" *In Proceedings Interspeech*, 2005

[16] K. Kinoshita , T. Nakatani and M. Miyoshi, "Spectral Subtraction Steered By Multi-step Forward Linear Prediction For Single Channel Speech Dereverberation" *In Proceedings IEEE Int. Conf. Acoust., Speech, Signal Proc. ICASSP*, 2006

[17] R. Gomez, J. Even, H. Saruwatari, and K. Shikano , "Distant-talking Robust Speech Recognition Using Late Reflection Components of Room Impulse Response" *In Proceedings IEEE Int. Conf. Acoust., Speech, Signal Proc. ICASSP*, 2008

[18] R. Gomez, J. Even, H. Saruwatari, and K. Shikano, "Fast Dereverberation for Hands-Free Speech Recognition" *IEEE Workshop HSCMA*, 2008

[19] H. Kuttruff, "Room Acoustics" *Spon Press*, 2000

[20] P. Naylor and N. Gaubitch, "Speech Dereverberation" *In Proceedings IWAENC*, 2005

[21] Y. Huang, J. Benesty, and J. Chen, "Speech acquisition and enhancement in a reverberant, cocktail-party-like environment" *In Proceedings IEEE Int. Conf. Acoust., Speech, Signal Proc. ICASSP*, 2008

[22] G. Gannot and M. Moonen, "Subspace Methods for Multimicrophone Speech Dereverberation" *In Proceedings Eurasip Journal on Applied Signal Processing*, vol. E80-A pp 804-808, 1997

[23] T. Hikichi, M. Delcroix, and M. Miyoshi, "Inverse Filtering for Speech Dereverberation Less Sensitive to Noise and Room Transfer Function Fluctuations" *In Proceedings Eurasip Journal on Advances in Signal Processing*, vol. 2007

[24] H. Attias, J. Platt, A. Acero, and L. Deng, "Speech Denoising and Dereverberation Using Probabilistic Models" *In Proceedings MIT Press In Advances in Neural Information Processing Systems 13*, 2001

[25] T. Nakatani, B-H. Juang, T. Yoshioka, K. Kinoshita, M. Delcroix, and M. Miyoshi, "Speech Dereverberation Based on Maximum-Likelihood Estimation with Time-Varying Gaussian Source Model" *In Proceedings IEEE Trans. on Audio, Speech, and Lang. Proc.*, Vol. 16, No. 8, 2008

[26] R. Gomez, T. Kawahara, K. Nakamura and K. Nakadai, "Robust hands-free Automatic Speech Recognition for human-machine interaction" *In Proceedings IEEE Humanoids*, 2010

[27] H. Sawada *et al.*,"Polar coordinate based nonlinear function for frequency-domain blind source separation," in *Proc. of ICASSP 2002*, 2002

[28] H. Nakajima, K. Nakadai, Y. Hasegawa and H. Tsujino, "Adaptive Step-size Parameter Control for real World Blind Source Separation" *In Proceedings IEEE Int. Conf. Acoust., Speech, Signal Proc. ICASSP*, 2008

[29] Akinobu Lee "JULIUS: A Free Continuous Speech Recognition Software" *www.sourceforge.jp* Kyoto University, Japan

[30] L.R.Rabiner and B. Gold, "Theory and Application of Digital Signal Processing" *Prentice Hall, Englewood Cliffs* 1975

[31] L.R.Rabiner and R.W. Scahefer, "Digital Processing of Speech Signals" *Prentice Hall, Englewood Cliffs* 1978

[32] L.R.Rabiner and B.H. Juang , "Fundamentals of Speech Recognition" *Prentice Hall, Englewood Cliffs* 1993

[33] C.H. Lee, L.R. Rabiner, R. Pieraccini and J.G. Wilpon "Acoustic Modelling for Large Vocabulary Speech Recognition" *In Proceedings Computer Speech and Language*, 1990

[34] T. Cincarek, H. Kawanami, H. Saruwatari, and K. Shikano,"Development and portability of ASR and Q and A modules for real-environment speech-oriented guidance systems" *In Proceedings IEEE Automatic Speech Recognition and Understanding ASRU*, 2007

[35] S. Takeuchi, T. Cincarek, H. Kawanami, H. Saruwatari, and K. Shikano,"Question and answer database optimization using speech recognition results" *In Proceedings Interspeech*, 2008

[36] Y. Suzuki, F. Asano, H.-Y. Kim, and T. Sone, "An optimum computer-generated pulse signal suitable for the measurement of very long impulse responses" *Journal Acoustical Society of America*, 1995

[37] H.-G. Hirsch and H. Finster, "A new approach for the adaptation of HMMs to reverberation and background noise" *In Proceeding Speech Communication*, pp 244-263, 2008

[38] R. Gomez, K. Nakamura and K. Nakadai, "Hands-free Human-Robot Communication Robust to Speaker's Radial Position" *In Proceeding IEEE International Conference on Robots and Automation ICRA*, 2013