Proceedings of the 35th Chinese Control Conference
July 27-29, 2016, Chengdu, China

Intelligent Speech Control System for Human-Robot Interaction

Xiaomei Liu, Shuzhi Sam Ge, Rui Jiang, and Cher-Hiang Goh

Abstract— **Accurately extracting subjective contents of speech signals and applying it on controlling robots remain to this day a challenging task as well as an insistent demand in human-robot interaction (HRI). A simple classification of human's intentions may limit the development of robots' natural reactions to users. Additionally, there should be a control system that can understand and translate human's intentions into control inputs. This paper proposes an intelligent speech control system for HRI. The objective is to understand human' s speech commands via recognizing, quantifying audio signals and translating speech inputs into control inputs. Aiming at this purpose, three main parts for the system are designed: a speech recognition system, a speech measurement system and a control system. Specifically, Mel Frequency Cepstral Coefficient (MFCC) and Dynamic Time Warping (DTW) techniques are utilized to recognize isolate speech commands. An energy-based feature and novelly proposed spectrum-based features are introduced to represent subjective contents of speech signals followed by Random Forest (RF) as a regressor. Several control schemes are utilized to translate quantified speech signals into control inputs. Simulation results illustrate the performance of the proposed system and the robot adaptive control system outperforms other control methods on effectiveness and controllability. The improved spectrum-based features demonstrate the capacity to extract subjective information of signals.**

I. Introduction

The field of HRI has been widely studied over the past few yeas because of people's increasing demands of such robots and the increasing complexity of robots including computational capabilities and perception capabilities, etc. A variety of perception signals are applied to realized HRI, such as speech [1], image [2] and brain-generated signals [3]. Among these, speech signals should be the most natural and convenient way benefiting from low requirements of devices and easy perception.

The recognition of objective contents, or generally called as speech recognition (SR), has been intensively studied in the previous research work and excellent achievements have been obtained. The typical techniques includes applying the expectation-maximization (EM) algorithm for training hidden Markov models (HMMs) [4]; modelling audio signals by concatenating Mel-frequency cepstral coefficients (MFCCs) [5] or perceptual linear predictive coefficients (PLPs) [6] computed from the raw waveform and their first- and second-order temporal differences [7]; artificial neural networks (ANN) with a single layer of nonlinear hidden units to predict HMM states from windows of acoustic coefficients [8]; and deep neural networks (DNNs) containing many layers of nonlinear hidden units and a large output layer [9].

Despite great progress made in SR, we are still far from having natural interactions between human and robot because robot can not understand exactly the subjective information of the speaker's speech. The recognition of subjective contents plays a meaningful role in HRI. Regrettably, the recognition of human's speech emotion has been almost exclusively studied in recent years [10][11], but it is far from being desired. Especially, few researchers have considered to quantify subjective contents of speech signals. This issue is significant since the detection of human's intentions is usually one step of HRI. The outcome of this step will be further utilized as the input of the subsequent processes to promote the robot to have a more proper response. A simple classification mechanism is unable to meet human's demands for high intelligent robots.

Besides, perception results from speech and brain signals are generally treated as inputs for knowledge-based expert system [12][13] to decide motions of robot. One of important factors resulting in this block is the challenge of the quantitative analysis for speech signals. Another condition accelerating the improvement of robotic intelligence is a control system that can accurately map perception results into control inputs. To achieve this, many challenges have to be solved, such as users' diversity and environmental uncertainties.

Therefore, this study proposes an intelligent speech control system that can better understand human's speech commands and accurately translate commands into continuous control inputs for the robot. The main contributions are as below:

- A speech measurement system is designed for the quantitative analysis of acoustic signals based on an energy-based feature or novelly proposed spectrum-based features followed by RF as a regressor. Spectrum-based features can model the natural pattern by which people express their intentions and has the potential to be utilized on emotion detection.
- A robot adaptive control system is proposed to obtain control inputs from quantified speech signals. Simulation results show the algorithm is more efficient and controllable compared to other control schemes. Furthermore, it can eliminate influences of environment and human subjective factors.
- The combination of above two systems translate speech signals into continuous control inputs. This technique can be used in many other applications, such as smart home systems and unmanned ground vehicles (UGVs).

Xiaomei Liu, Shuzhi Sam Ge, Rui Jiang and Cher-Hiang Goh is with the Department of Electrical and Computer Engineering, National University of Singapore, Singapore 117576 xiaomeiliu@nus.edu.sg, samge@nus.edu.sg, rui_jiang@u.nus.edu, elegch@nus.edu.sg

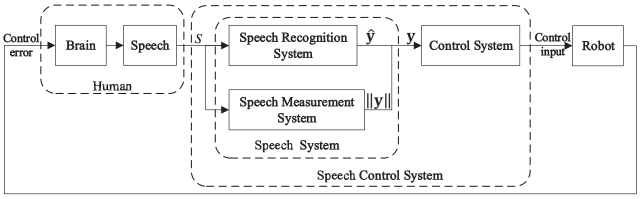


Fig. 1: Speech control system diagram

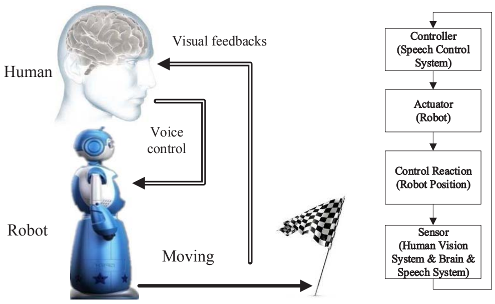

Human
Visual feedbacks
Voice control
Robot
Moving
Controller (Speech Control System)
Actuator (Robot)
Control Reaction (Robot Position)
Sensor (Human Vision System & Brain & Speech System)

Fig. 2: Speech control system working principle

II. Speech Control System

A speech control system is developed that can better understand human's speech commands (including recognizing and measuring), and translate them into control inputs. The overall system diagram is shown in Fig. 1. Human visual system observes control errors, and inputs observed errors into the brain. After that, the brain will control the human speech system to express control intentions as inputs of the speech control system, which includes three main parts. The first part is a speech recognition system, which executes the qualitative analysis, that is, recognizes objective contents of speech; the second part is a speech measurement system which conducts the quantitative analysis, that is, measures subjective contents of speech; the third part is a control system which translates fused results of the speech system into control inputs. The speech recognition and measurement system are collectively called as the speech system. These three parts will be explained in detail in subsequent sections.

The speech control system can detect control intentions sent by brain, expressed by human speech system and translate speech signals into control inputs (see Fig. 2). Considering the robot as an actuator, and the brain as a controller, which provides an initial control input based on vision guidance of the relative position of the robot and the target. After that, in the process of approaching the target, the brain will adaptively adjust speech inputs so that the robot could reach the target fast and smoothly. The goal of the system control system is to accurately detect speech inputs and translate them into control inputs.

III. Speech Recognition System

Assume $\mathbf{u} = [u_1, ..., u_n]$ to represent the control input of the robot, where n is the dimensionality of the control input, and u_i denotes the i^{th} dimension of $\mathbf{u}$. Every two isolated speech commands are utilized to be mapped into one dimension of the control input as $\mathcal{Y} = \{y_1^+, y_1^-, ..., y_n^+, y_n^-\}$, where y_i^+ and y_i^- are corresponding to positive and negative directions of u_i. Therefore, to control a system with n dimensions of the control input, there will be totally $2n$ classes of speech commands that needed to be recognized.

A. Qualitative Feature Extraction

In the isolated command recognition system, features extracted must eliminate the influence from environment and human subjective factors, including emotion and health conditions, etc. Only objective contents should be reserved. The extraction of the best parametric representation of acoustic signals is an important task to produce a better recognition performance. The accuracy of this phase is crucial for the next phase since it directly decide the sign of control inputs. Here, MFCC is applied to represent the acoustic input [5]. The overall process of the MFCC is shown in Fig. 3 marked in black.

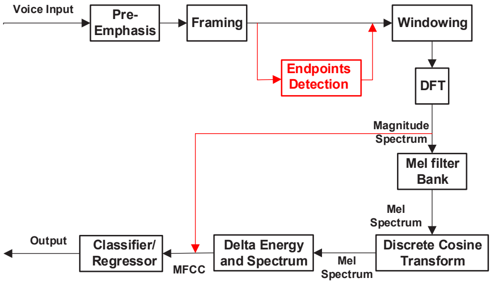


Fig. 3: Comparison of MFCC and proposed spectrum-based features

B. Speech Recognition by Template Matching

In this section, a map function $f_1 : s \to \hat{\mathbf{y}}$ is going to be built, where s is an audio signal, and $\hat{\mathbf{y}} \in \mathbb{R}^n$ is an unit vector. Denote y_i as the i^{th} dimension of $\hat{\mathbf{y}}$, and if $s \in y_i^+$, then $y_i = 1$; if $s \in y_i^-$, then $y_i = -1$; if $i' \neq i$, $y_{i'} = 0$. Furthermore, define u_0 as a positive value representing a control value translated from speech inputs, if $y_i(j) = 1$ (y_i at time j), then $u_i = u_0$; if $y_i(j) = -1$, then $u_i = -u_0$. It is worth noting that only one $y_i = \pm 1, i = 1, .., n$ at the moment of there is a speech input. In the next section, we will discuss how to obtain $\|\mathbf{y}\|$, which is the measurement of $\hat{\mathbf{y}}$.

DTW algorithm is based on Dynamic Programming techniques as described in [14]. This algorithm aims at measuring similarity between two time series which may vary in time or speed. In this paper, we apply DTW [15] to calculate the distance vectors of MFCC series for measuring the similarity between audio signals, and Euclidean distance to calculate distance vectors of audio feature series.

After distances between the current speech feature series and each speech feature series from training templates are obtained, the target of the one in training templates with the maximum similarity will be assigned to the testing sample.

IV. Speech Measurement System

There are two broad types of information in speech. The semantic part of the speech carries objective information insofar that the utterances are made according to the rules of pronunciation of the language. Subjective information, on the other hand, refers to the implicit messages such as the emotional state of the speaker or the control intention, which will be studied in this section.

A. Quantitative Feature Extraction

A set of features that can precisely represent control intentions are extremely significant to the accuracy of quantified outcomes. Two possible candidate features are studied as below.

1) *Energy-based feature*

Volume indicates the speech intensity, which could be represented as the amplitude of signal in each frame. It is the most direct way to express control intentions. The essential steps extracting energy-based features are listed as below.

Step 1: Framing

The process of segmenting the speech samples obtained from analog to digital conversion (ADC) into a small frame with the length within the range of 20 to 40 msec. The voice signal is divided into frames of N samples. Adjacent frames are being separated by M $(M < N)$. In this paper, we choose $M = 128$ and $N = 256$.

Step 2: Endpoints detection

Endpoint detection of speech signal is a step directly affecting the accuracy of quantitative outcome. Here, a dual-threshold speech endpoint detection algorithm with the use of short-term energy and short-term average zero-crossing rate is applied [16].

Step 3: Average active energy

The average energy between two endpoints will be calculated as quantified results of the intensity of control intentions to be further mapped as control inputs for the robot.

Several drawbacks of this method exist: (i) not all people express their control intensions through changing volume; (ii) the sound source must be fixed; (iii) it is sensitive to background noise. However, the advantages are (i) the model training process is not needed; (ii) although changing volume may not be a natural way to express control intentions, it is the most direct and easily controlled way under the condition that users know this rule.

2) *Spectrum-based feature*

Step 1: Framing (the same as above)

Step 2: Endpoints detection (the same as above)

Step 3: Hamming windowing

The Hamming window equation is given as $Y(n) = X(n) \times W(n)$, where $Y(n)$ is the output signal, $X(n)$ is the input signal and $W(n)$ is the window defined as $W(n) = 0.54 - 0.46\cos\left(2\pi n/\left(N-1\right)\right), 0 \leq n \leq N-1$.

Step 4: Fast Fourier Transform

$$Y(\omega) = \mathrm{FFT}[H(n) * X(n)] = H(\omega) * X(\omega) \quad (1)$$

where $X(\omega)$, $H(\omega)$ and $Y(\omega)$ are the Fourier Transform of $X(n)$, $H(n)$ and $Y(n)$ respectively.

The overall progress of extracting spectrum-based features is shown in Fig. 3 marked in red. Compared with commonly used spectral features, e.g., LPCC, MFCC and LFPC, etc., powers of the spectrum are directly taken as the input of subsequent classifier or regressor, rather than being mapped using a set of given filters and taken DCT to obtain powers sequence in time domain. The motivation of this improvement is that diverse users may emphasis different frequencies to express their control intentions, and the subsequent regressor is able to learn this pattern.

The pros and cons of spectrum-based features are nearly opposite to the ones of the energy-based feature. For its superiority, (i) it is robust to background noise; (ii) being able to learning a natural pattern by which people express control intensions; (iii) the sound source may be mobile. For shortages, (i) the training process is required, that is, a large amount of training data should be provided; (ii) the training target is man-made given, so the calibration error would be another error source for control; (iii) although this method can learn a natural control pattern, the pattern may be adapted to the speaker's emotion, environment and time, and it is difficult to be expressed deliberately.

B. Speech Measurer

In this section, a map function $f_2 : s \rightarrow \|\mathbf{y}\|$ is going to be built. That is to decide after a set of features are extracted that may represent the intensity of the user's control intentions, how to map them into a control magnitude. For the energy-based feature which is a scalar, only a constant coefficient is needed to linearly map the extracted speech feature into a control input. For spectrum-based features which is a feature vector, a step of regression is required.

Therefore, for spectrum-based features, a regressor named *Random Forest* is applied here [17]. It is a type of ensemble classification that uses decision tree as the base classifier. RF is chosen as the regressor for several main reasons: (i) high speed (ii) high accuracy (iii) capability to evaluate the importance of each feature variable (iv) being able to handle a feature vector with high dimension, which means the feature selection is not required. Especially, the high computation speed is extremely crucial to guarantee the real-time performance of the speech control system.

V. Control System

Several control schemes will be described in this section and compared in next section such that the quantitative outcomes of the speech system can be translated into continuous control inputs.

A. System Dynamics

Consider a linear system dynamics as

$$\begin{cases} x_1(j+1)-x_1(j)=v(j)\cos\left(\psi(j)\right) \\ x_2(j+1)-x_2(j)=v(j)\sin\left(\psi(j)\right) \\ \psi(j+1)-\psi(j)=\omega(j) \end{cases} \tag{2}$$

where x_1 and x_2 refers to position coordinates, and ψ refers to the motion direction. Denote the control input as $u(j)=\begin{bmatrix} v(j) & \omega(j) \end{bmatrix}^{\mathrm{T}}$.

In this paper, as the environment is assumed to be unknown to the robot, control errors cannot be measured directly by the robot, but can be observed by the human. That means the robot can only sense the environment by outputs of the audio sensor. This assumption may be too incredible in realistic; however, it is to motivate the development of a speech-based HRI system used as the auxiliary correction system in real applications. `otherwise`

B. Controller Design

A common way to realize this mission is that every time there is a speech command, robot turns a fixed angle, which can be represented as

$$\omega(j)=c \tag{3}$$

where c is a positive constant. The problem is that a small angle c may render a slow control speed; on the contrary, if c is too large, it may be very difficult to reach the desired direction exactly.

Alternatively, to enable the robot better understand human's commands, control laws are going to be designed for the speech control system under conditions that control errors can only be perceived by the human, while the robot can indirectly measure control errors through speech signals delivered by the human. Denote

$$f: s \rightarrow \mathbf{y} \tag{4}$$

as fused outputs of the speech system, including the speech recognition system ($f_1: s \rightarrow \hat{\mathbf{y}}$) and the speech measurement system ($f_2: s \rightarrow \|\mathbf{y}\|$). Totally consider environmental noise and system noise as ξ, and simply write $\|\mathbf{y}\|$ as $\hat{y}$. The estimation progress of the speech measurement system can be marked as $\hat{y}=f_2(s)+\xi$. In this section, the noise is ignored such that $\xi=0$. The progress function can be rewritten as $\hat{y}=f_2(s)$. The human brain map function from observed control errors to control intensions is defined as $f_3: \psi_e \rightarrow y$, which is unknown actually. The main purpose of this section is to find an approximation of f_3, and $\hat{y}$ is the estimation of y. We assume $\hat{y}=y$ here, which means the outcome of speech system is assumed to be equal to the human control intention.

Meanwhile, we assume the linear uncoupling relationship between ψ_e and y, which simplifies map function $y=f_3(\psi_e(j))$ into $y(j)=a^*\psi_e(j)$ at time j, where a^* is an unknown positive constant. We denote the inverse of a^* as g^*, that is $g^*={a^*}^{-1}$. Similarly, g^* is an unknown positive constant. Then we have $\psi_e(j)=g^*y(j)$.

A. Human Adaptive Control System

The control law $\omega(j)$ is chosen to be

$$\omega(j)=k_y y(j) \tag{5}$$

where $k_y>0$ is a constant. In this method, a constant coefficient k_y is given to map the measurement result to control direction error, then human brain will gradually figure out this coefficient after several attempts. Therefore, we call this method as "*Human Adaptive Control*". However, the aim of designing the speech control system is that robot could better understand human's intentions rather than human trying to adapt to the system. Therefore, the robot adaptive control system is designed for the sake of a more natural HRI performance, which will be introduced in the following section.

B. Robot Adaptive Control System

In this method, the control task is divided into two phases: adaptive phase and convergence phase. In the adaptive phase, the focus is to find an approximation of g^*; while in the second phase, the control law is implemented with g^* obtained in the first phase to perform tasks, such as target tracking and obstacle avoidance. Firstly, the control law $\omega(j)$ is designed to be

$$\omega(j)=k_\omega\psi_e(j)=k_\omega g^* y(j) \tag{6}$$

where $k_\omega>0$ is a constant. However, as g^* is unknown resulting in that the control law cannot be utilized directly, so Eq. (6) is rewritten as

$$\omega(j)=k_\omega g(j)y(j) \tag{7}$$

with a random initialization of $g(j)$, when there is a speech input at time j. The adaptive law for $g(j)$ is designed as

$$g(j+1)=g(j)+\frac{k_g}{k_\omega}\frac{y(j+1)-(1-k_\omega)\,y(j)}{y(j)} \tag{8}$$

where k_g is a positive constant. After the first phase, an estimated desired $\hat{g}^*$ can be obtained, which will be used for the task in the next phase. In the second phase, the control law is chosen to be

$$\omega(j)=k_{\mathrm{t}}k_\omega\hat{g}^* y(j) \tag{9}$$

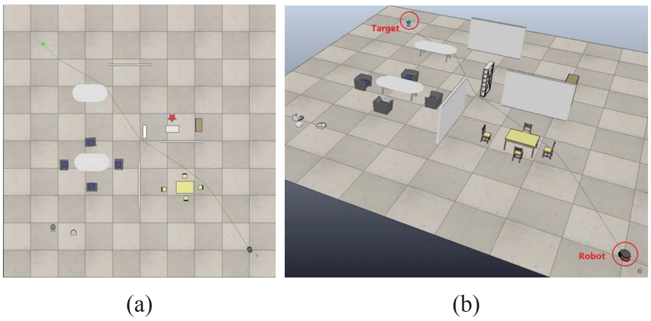
where $k_t>1$ is a constant introduced to reduce the influence of time delay in the system due to communication delay and computation cost on the speech system.

There is another control scheme, which is to implement Eq.(7) & Eq.(8) for the whole task. The merit is that the possible range of control inputs is broader and the sensitivity is improved; nevertheless, the degree of controllability is weaken.

C. Acceleration strategy

A small value of velocity $v(j)$ may be beneficial to the direction adjustment, while a large value of $v(j)$ may be beneficial to reducing task time cost. To balance two factors, a relative small value of v_0 is set in the beginning and a window function $W(j)$ with a length of l is introduced that

$$W(j)=\begin{cases} 1 & \text{if } 0\leq j\leq l-1 \\ 0 & \text{otherwise} \end{cases} \tag{10}$$



(a) (b)

Fig. 4: Simulation environment

then the acceleration strategy is designed as

$$v(j) = \begin{cases} k_v v(j-1) & \text{if } \sum_{i=0}^{j-1} y(j-i)W(i) \leq \varsigma \\ v_0 & \text{otherwise} \end{cases} \tag{11}$$

where $k_v > 1$ is a constant set for acceleration, and ς is a very small positive value. The reason why we do not set it as zero is to avoid the influence of noise.

VI. Simulation studies

A. System Setup

The speech control system is implemented in *Python*, while *V-REP* is chosen as the robot simulator. The simulation environment is built as a simple indoor environment, as shown in Fig. 4. We choose *Pioneer 3DX* as the simulation robot. The robot is initialized at a fixed point, the control objective is to reach the point marked by the green plant. The task space around the initial position is relative spacious such that g is able to converge to the desired value g^*, while the space for the second phase is full of obstacles to test the effectiveness of the proposed algorithm.

An optimal path generated by the path planning module in *Python* is plotted in Fig. 4. It is treated as a reference for the results analysis.

B. Training data

For the model training of spectrum-based features, five targets (0.2, 0.4,...,1) are assigned as outcomes of RF. Training samples are collected round by round. Each round is required to express the control intensity in order from 0.2 to 1. One subject took two rounds of data collection in each time slot, and there are three time slots arranged every day, respectively in the morning, afternoon and evening. It took 5 days to complete data collection. Thus, there are 30 training samples for each targets and 150 training samples in total.

C. Results & analysis

5 groups of trials are designed with the energy-based feature while using diverse control schemes to be compared. Time cost and the number of total speech commands are chosen to indicate the performance of approaches. The detailed control schemes for each group are listed as below:

- *Group 1*: Eq.(3) & Eq.(11);
- *Group 2*: Eq.(5) & Eq.(11);
- *Group 3*: Eq.(7) & Eq.(8) & Eq.(11);
- *Group 4*: Eq.(7) & Eq.(8) & Eq.(11) in the first phase and Eq.(9) & Eq.(11) in the second phase;
- *Group 5*: Eq.(7) & Eq.(8) in the first phase and Eq.(9) in the second phase with $v(j) = v_0$.

Trials for each group are repeated for 10 times. Especially, parameters involved like c in Eq.(3) and k_y in Eq.(5) are properly chosen according to common practices. This is to basically guarantee equal conditions for each group and the comparability of results (see Table.I). Trajectories of one of the trials for Groups 1-4 are shown in Fig.5, and the optimal path is also marked as the reference.

TABLE I: Results comparison of Groups 1-6

Grp No.	factors	Average	MIN	MAX	VAR
1	time (s) commands number	172.3 17.9	158 13	198 22	137.81 9.09
2	time (s) commands number	160.4 15.3	137 12	176 20	116.04 4.81
3	time (s) commands number	144.7 13.1	125 10	154 17	56.81 5.29
4	time (s) commands number	138.8 12.2	127 9	142 14	17.56 1.76
5	time (s) commands number	193.4 12.2	192 9	196 15	2.24 4.16
6	time (s) commands number	130.5 11.7	112 7	141 15	67.45 4.81

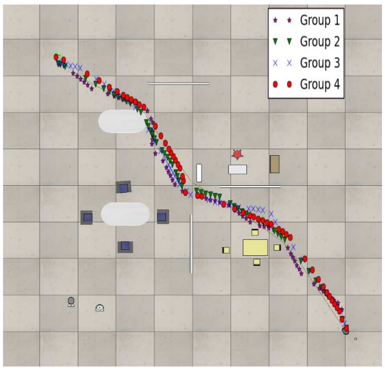


Fig. 5: Trajectories of Groups 1-4

Above results show that *Group 4* has the best performance no matter on the time cost or on the number of commands. Except *Group 5*, the results of *Group 4* have the minimum variance. This means that the algorithm for *Group 4* is less user-related, and will be less influenced by external factors, such as time, user's emotion and environment. This is because the algorithm is to make the system to adapt to the human.

By contrast, algorithms for *Group 2* require user's self-adaptation to the system. Thus, any factors influencing the user will easily influence the performance. Results of *Group 1* are worst, where the reason has been stated in the beginning of *Section 5*. Outcomes for *Group 3* are much more satisfactory compared to the ones of *Groups 1-2*. The problem is in the second half of the task, the robot's motion direction has to be frequently changed to avoid collisions. However, these is no convergence phase for *Group 3*, and $g(j)$ will be adjusted for the whole task so the sensitivity of the method is relatively high, which influences the controllability of the robot.

Comparing results of *Group 4* with the ones of *Group 5*, the only difference of results is the time cost, which is due to the usage of *acceleration strategy*. Furthermore, a smaller variance of the time cost for *Group 5* results from the robot with a small uniform velocity is more easily controlled.

Fig.5 indicates that trajectories generated by methods of Groups 1-4 are all quite similar and close to the reference path. Reconsider results in Table.I, a crucial factor determining the time cost could be found, which is the total number of commands required. The time cost is proportional to the commands number. A better control scheme enables the robot achieve desired motion direction faster, thus less commands are required leading to less time cost.

Next, to prove the effectiveness of proposed spectrum-based features, another *Group 6* is designed with the similar conditions as *Group 4* but is carried out using spectrum-based features. The number of trees for RF is set to be 500. 10-fold cross validation is applied to select the RF model with the best generalization ability. Results (see the last row of Table.I) show the average time cost and the number of speech commands required is less. Thus, this to a certain extent proves that human's natural speech control pattern is learned by the proposed approach. However, the variance is much higher as spectrum-based features are not as easily controlled by inexpert users as the energy-based feature.

To further show the effectiveness of the proposed spectrum-based features in extracting subjective information of speech signals, we compare outcomes of the proposed spectrum-based features and MFCC by calculating mean square error (MSE) and correlation coefficient (CC) of targets and outputs of RF as indexes. CC is obtained here to illustrate the effectiveness because the training data is collected round by round. There may be bias between data in different round, but in contrast, positive correlation for the data in one round should be more meaningful. Results are shown in Table.II, which indicate the proposed approach has better performance than MFCC.

TABLE II: Results comparison of MFCC & proposed spectrum-based features

Method	Command	MSE	CC
MFCC	Left	5.63×10^{-3}	0.975
	Right	9.28×10^{-3}	0.964
Our method	Left	2.18×10^{-3}	0.989
	Right	2.20×10^{-3}	0.991

VII. Conclusion

In this paper, a speech control system for the application on HRI has been proposed. Three main parts have been designed such that the robot can better understand human' s speech commands, including the speech recognition system, the speech measurement system and the control system. The system is able to detect human's intentions via audio perceptions and translate speech signals into continuous inputs for motion control of robot. Finally, simulation results have been presented, compared and analyzed to illustrate the performance of proposed system. The proposed robot adaptive control system with acceleration strategies has shown the highest effectiveness and controllability. Meanwhile, it has been proved that the improved spectrum-based features demonstrate the potential in detecting subjective information of speech signals. This method can be further developed for emotion recognition.

References

[1] Amin Atrash, Robert Kaplow, Julien Villemure, Robert West, Hiba Yamani, and Joelle Pineau. Development and validation of a robust speech interface for improved human-robot interaction. *International Journal of Social Robotics*, 1(4):345–356, 2009.

[2] Y Yang, Shuzhi Sam Ge, Tong Heng Lee, and C Wang. Facial expression recognition and tracking for intelligent human-robot interaction. *Intelligent Service Robotics*, 1(2):143–157, 2008.

[3] Xavier Perrin, Ricardo Chavarriaga, Francis Colas, Roland Siegwart, and José del R Millán. Brain-coupled interaction for semi-autonomous navigation of an assistive robot. *Robotics and Autonomous Systems*, 58(12):1246–1255, 2010.

[4] Sadaoki Furui. Digital speech processing, synthesis, and recognition (revised and expanded). *Digital Speech Processing, Synthesis, and Recognition (Second Edition, Revised and Expanded)*, 2000.

[5] Steven B Davis and Paul Mermelstein. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, 28(4):357–366, 1980.

[6] Hynek Hermansky. Perceptual linear predictive (plp) analysis of speech. *the Journal of the Acoustical Society of America*, 87(4):1738–1752, 1990.

[7] Sadaoki Furui. Cepstral analysis technique for automatic speaker verification. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, 29(2):254–272, 1981.

[8] Herve A Bourlard and Nelson Morgan. *Connectionist speech recognition: a hybrid approach*, volume 247. Springer Science & Business Media, 2012.

[9] Geoffrey Hinton, Li Deng, Dong Yu, George E Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara N Sainath, et al. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *Signal Processing Magazine, IEEE*, 29(6):82–97, 2012.

[10] Jia-Ching Wang, Yu-Hao Chin, Bo-Wei Chen, Chang-Hong Lin, and Chung-Hsien Wu. Speech emotion verification using emotion variance modeling and discriminant scale-frequency maps. *Audio, Speech, and Language Processing, IEEE/ACM Transactions on*, 23(10):1552–1562, 2015.

[11] Norhaslinda Kamaruddin, Abdul Rahman, Abdul Wahab, and Nor Sakinah Abdullah. Speech emotion identification analysis based on different spectral feature extraction methods. In *Information and Communication Technology for The Muslim World (ICT4M), 2014 The 5th International Conference on*, pages 1–5. IEEE, 2014.

[12] Shi-An Chen, Chih-Hao Chen, Jheng-Wei Lin, Li-Wei Ko, and Chin-Teng Lin. Gaming controlling via brain-computer interface using multiple physiological signals. In *Systems, Man and Cybernetics (SMC), 2014 IEEE International Conference on*, pages 3156–3159. IEEE, 2014.

[13] Bjorn Schuller, Gerhard Rigoll, Salman Can, and Hubertus Feussner. Emotion sensitive speech control for human-robot interaction in minimal invasive surgery. In *Robot and Human Interactive Communication, 2008. RO-MAN 2008. The 17th IEEE International Symposium on*, pages 453–458. IEEE, 2008.

[14] Stan Salvador and Philip Chan. Toward accurate dynamic time warping in linear time and space. *Intelligent Data Analysis*, 11(5):561–580, 2007.

[15] Andres Marzal and Enrique Vidal. Computation of normalized edit distance and applications. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 15(9):926–932, 1993.

[16] Qiuyu Guo, Nan Li, and Guangrong Ji. A improved dual-threshold speech endpoint detection algorithm. In *Computer and Automation Engineering (ICCAE), 2010 The 2nd International Conference on*, volume 2, pages 123–126. IEEE, 2010.

[17] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.