Intelligent Speech Control System for Human-Robot Interaction

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.

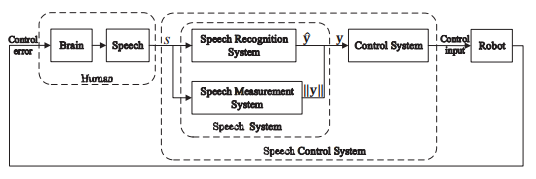


Fig. 1: Speech control system diagram

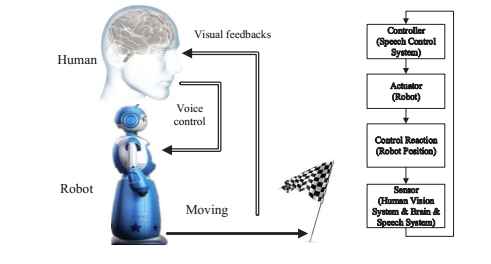
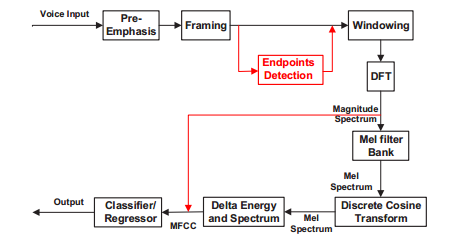


Fig. 2: Speech control system working principle

SPEECH RECOGNITION SYSTEM  
 Assume u = [u1, ..., un] to represent the control input of the robot, where n is the dimensionality of the control input, and ui denotes the ith dimension of u. Every two isolated speech commands are utilized to be mapped into one dimension of the control input as Y = {y1 +, y1 -, ..., yn +, yn -}, where yi + and yi - are corresponding to positive and negative directions of ui. 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 . 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 f1 : s → y ˆ is going to be built, where s is an audio signal, and y ˆ ∈ Rn is an unit vector. Denote yi as the ith dimension of y ˆ, and if s ∈ yi +, then yi = 1; if s ∈ yi -, then yi = -1; if i = i, yi = 0. Furthermore, define u0 as a positive value representing a control value translated from speech inputs, if yi(j) = 1 (yi at time j), then ui = u0; if yi(j) = -1, then ui = -u0. It is worth noting that only one yi = ±1, i = 1, .., n at the moment of there is a speech input. In the next section, we will discuss how to obtain \_x0005\_||y\_x0005\_||, which is the measurement of y ˆ.

DTW algorithm is based on Dynamic Programming techniques . This algorithm aims at measuring similarity between two time series which may vary in time or speed. In this paper, we apply DTW 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.

# 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 .  
**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)× 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 - 0.46cos(2πn/ (N - 1)) , 0 ≤ n ≤ N - 1.

**Step 4:** Fast Fourier Transform

Y (ω) = FFT[H(n) ∗ X(n)] = H(ω) ∗ X(ω) (1)

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

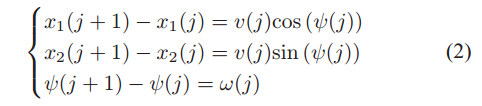
The overall progress of extracting spectrum-based features is shown in Fig. 3 marked in red. Compared with commonly used spectral features, e.g., LPCC, MFCC and LFPC, etc., powers of the spectrum are directly taken as the input of subsequent classifier or regressor, rather than being mapped using a set 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 manmade 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 f2 : s →||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 energybased 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. 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 realtime performance of the speech control system.

# CONTROL SYSTEM

Several control schemes will be described in this 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  
   
where x1 and x2 refers to position coordinates, and ψ refers to the motion direction. Denote the control input as u(j) = v(j) ω(j)\_x0007\_T.

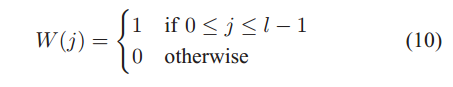
In this paper, as the environment is assumed to be unknown to the robot, control errors cannot be measured directly by the robot, but can be observed by the human. That means the robot can only sense the environment by outputs of the audio sensor. This assumption may be too incredible in realistic; however,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

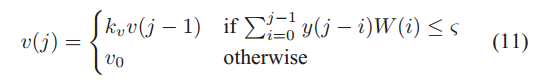
Alternatively, to enable the robot better understand human’s commands, control laws are going to be designed for the speech control system under conditions that control errors can only be perceived by the human, while the robot can indirectly measure control errors through speech signals delivered by the human. Denote  
 f : s → **y** (4)  
as fused outputs of the speech system, including the speech recognition system (f1 : s → y ˆ) and the speech measurement system (f2 : s → \_x0005\_y). Totally consider environmental noise and system noise as ξ, and simply write y\_x0005\_ as y ˆ. The estimation progress of the speech measurement system can be marked as y ˆ = f2(s) + ξ. In this section, the noise is ignored such that ξ = 0. The progress function can be rewritten as y ˆ = f2(s). The human brain map function from observed control errors to control intensions is de¿ned as f3 : ψe → y, which is unknown actually. The main purpose of this section is to ¿nd an approximation of f3, and y ˆ is the estimation of y. We assume y ˆ = 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 ψe and y, which simplifies map function y = f3(ψe(j)) into y(j) = a∗ψe(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 ψe(j) = g∗y(j).

Human Adaptive Control System .The control law ω(j) is chosen to be

ω(j) = kyy(j) (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”.

*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 v0 is set in the beginning and a window function W(j) with a length of l is introduced that  


then the acceleration strategy is designed as  
   
where kv > 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 inÀuence of noise.