题 目： 智能音箱语音控制系统的设计与实现

学院：＿＿＿＿＿＿专业：＿＿＿＿＿＿学生姓名：＿＿＿＿＿＿学号：＿＿＿＿＿＿

外文文献

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.

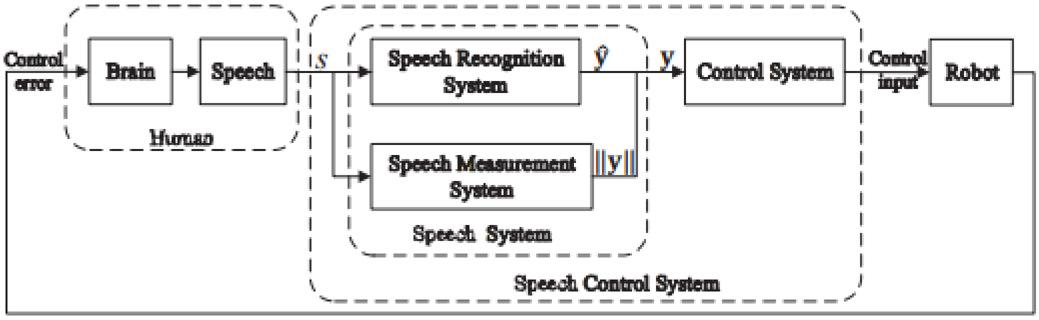

Fig. 1: Speech control system diagram

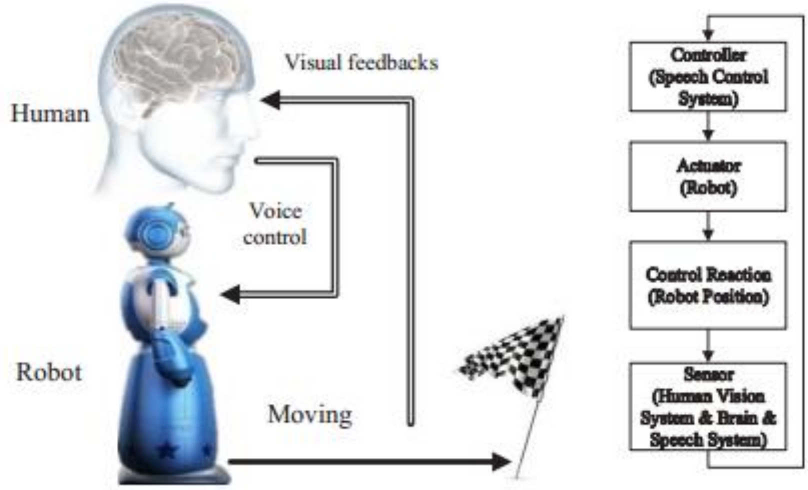
Human
Visual feedbacks
Voice control
Robot
Moving
Controller (Speech Control System)
Actuator (Robot)
Control Reaction (Robot Position)
Sensor (Human Vision System & Brain & Speech System)

Fig. 2: Speech control system working principle

SPEECH RECOGNITION SYSTEM

Assume $u=[u_1,\ldots,u_n]$ 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 u. Every two isolated speech commands are utilized to be mapped into one dimension of the control input as $Y=\{y_1^+, y_1^-, ..., y_n^+, y_n^-\}$, 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. The overall process of the MFCC is shown in Fig. 3 marked in black.

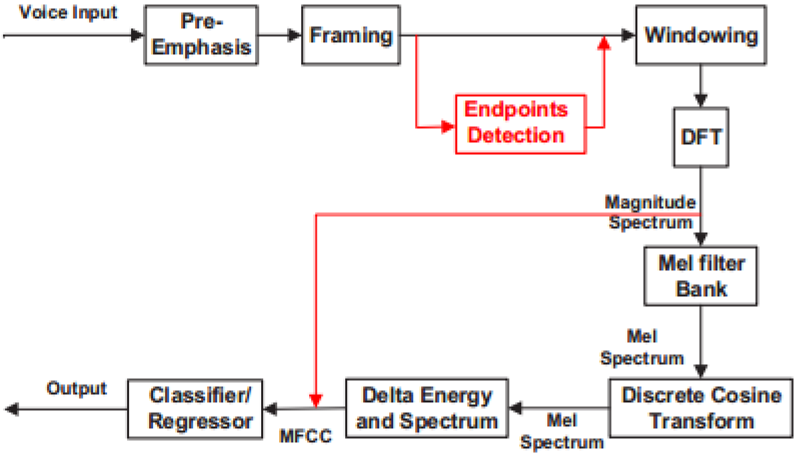

Fig. 3: Comparison of MFCC and proposed spectrum-based features

B. Speech Recognition by Template Matching

In this section, a map function $f_1 : s \to \hat{y}$ is going to be built, where s is an audio signal, and $\hat{y} \in \mathbb{R}^n$ is an unit vector. Denote y_i as the i^{th} ith dimension of $\hat{y}$, 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 $y_i = \pm 1$, $i = 1, \dots, n$ at the moment of there is a speech input. In the next section, we will discuss how to obtain $\|y\|$, which is the measurement of $\hat{y}$.

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 $W(n) = 0.54 - 0.46\cos(2\pi n/(N-1))$, $0 \le n \le N - 1$.

Step 4: Fast Fourier Transform

$$Y(w) = FFT\left[H(n) * X(n)\right] = H(w) * X(w) \qquad (1)$$

where $X(w)$, $H(w)$ and $Y(w)$ 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 $f_2 : s \rightarrow || y ||$ 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：

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

where x_1 and x_2 refers to position coordinates, and ψ refers to the motion direction. Denote the control input as $u(j) = [v(j)\omega(j)]^T$.

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

$$\omega(j)=c \qquad （3）$$

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

$$f: s \rightarrow y \qquad （4）$$

as fused outputs of the speech system, including the speech recognition system ($f_1: s \rightarrow \hat{y}$) and the speech measurement system ($f_2: s \rightarrow \| y \|$). Totally consider environmental noise and system noise as ξ, and simply write $\| y \|$ 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 ξ = 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 f3, 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 $y(j)=\alpha^* \psi_e(j)$ at time j, where α^* is an unknown positive constant. We denote the inverse of α^* as g^*, that is $g^*=\alpha^{*-1}$. Similarly, g^* is an unknown positive constant. Then we have $\psi_e(j)=g^* y(j)$.

Human Adaptive Control System. The control law $\omega(j)$ is chosen to be

$$\omega(j)=k_y y(j) \qquad （5）$$

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 v_0 is set in the beginning and a window function $W(j)$ with a length of l is introduced that

$$W(j)=\begin{cases}1 \ \mathit{if} \ 0\le j\le l-1 \\ 0 \ \mathit{otherwise}\end{cases} \qquad (6)$$

then the acceleration strategy is designed as：

$$v(j)=\begin{cases}k_v v(j-1) \ \mathit{if} \ \sum_{i=0}^{j=1} y(j-i)W(i)\le \varsigma \\ v_0 \ \mathit{otherwise}\end{cases} \qquad (7)$$

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 inÀuence of noise.

翻译

人机交互智能语音控制系统

语音控制系统用来更好地理解人类语音指令（包括识别和测量），并将其转换为控制输入。整体系统图如图1所示。人类视觉系统观察到控制错误，并输入观察到的错误进入大脑，之后，大脑会控制人类语音系统表达控制意图作为语音控制系统的输入。语音控制系统包括三个主要部分，第一部分是一个语音识别系统，它执行定性分析，即识别语音目的内容;第二部分是语音测量系统，进行定量分析，即测量语音主观内容;第三部分是控制系统把语音系统的融合结果转换成控制输入。语音识别和测量系统统一起称作语音系统，以下部分将详细说明这三部分。语音控制系统可以检测大脑发出由人类语音系统表达的控制意图，并将语音信号转换为控制输入（见图2)。考虑到机器人作为一个执行器，大脑作为控制器基于视觉引导的相对位置的机器人和目标提供初始控制输入，之后，在接近目标的过程中，大脑会自适应的调整语音输入，使机器人可以快速，顺利地理解目标。语音控制系统的目的是为了准确检测语音输入并将其转换为控制输入。

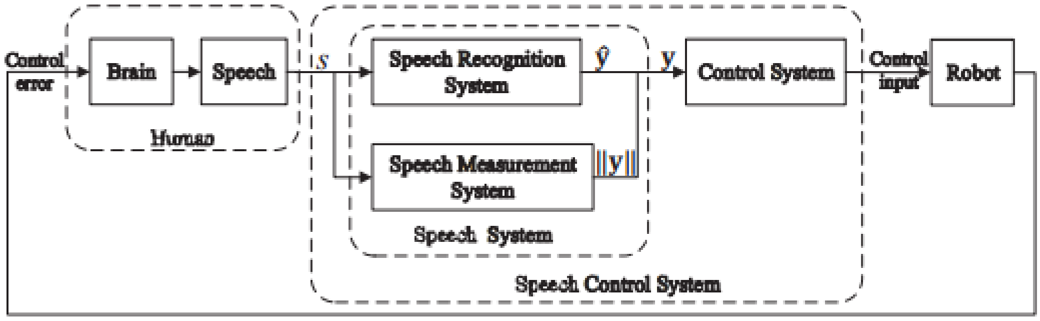

图1 语音控制系统图

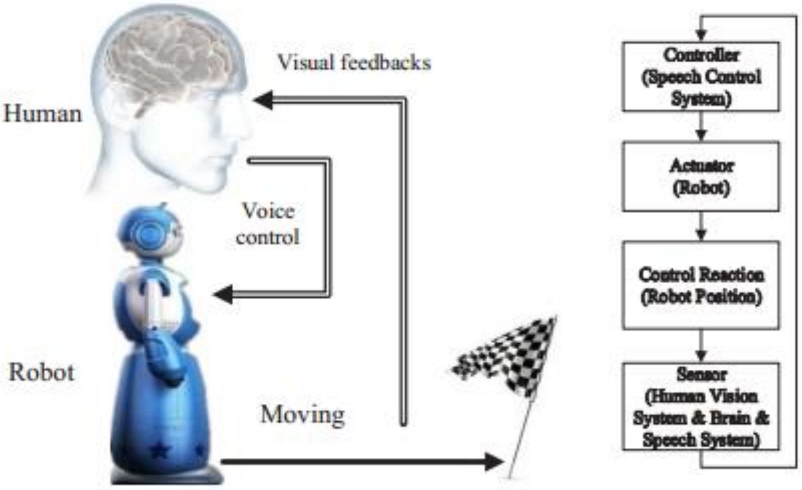
Human
Visual feedbacks
Voice control
Robot
Moving
Controller (Speech Control System)
Actuator (Robot)
Control Reaction (Robot Position)
Sensor (Human Vision System & Brain & Speech System)

图2 语音控制系统工作原理

语音识别系统

假设 u $=[u_1,\ldots,u_n]$表示机器人的控制输入，其中 n 是控制输入的维度，表示u_i的第 i 个维度。每两个孤立的语音命令被用来映射成一个控制输入的维度为$Y=\{y_1^+, y_1^-, ..., y_n^+, y_n^-\}$，其中$y_i^+$和$y_i^-$对应于$u_i$正和负方向。 因此，为了用控制输入的 n 个维度控制系统，这将产生 2n 个需要识别的语音指令类。

A.定性特征提取

在孤立的命令识别系统中，提取的特征必须消除环境和人类主观因素（包括情绪和健康状况）等的影响，只能保留客观内容。 提取声信号的最佳参数表示是产生更好的识别性能的重要任务。这一阶段的准确性对于下一阶段至关重要，因为它直接决定了控制输入的符号。 这里，MFCC 被应用于表示声输入。 MFCC 的整体过程如图 3 中标记为黑色的部分。

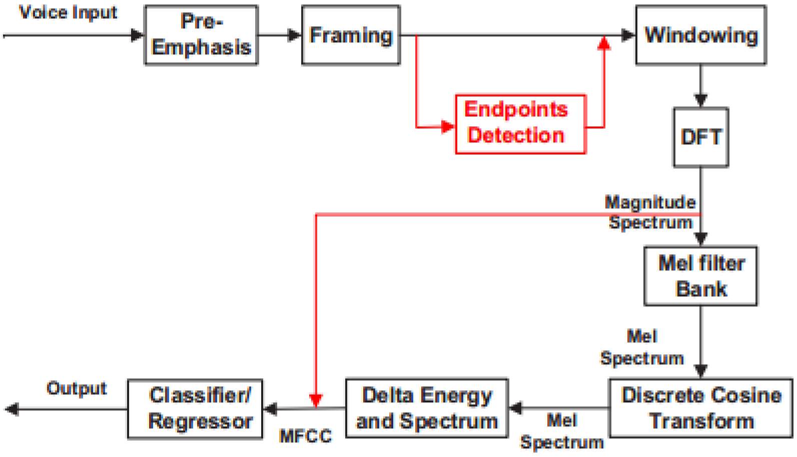

图 3 MFCC 和提出的基于频谱的特征的比较

B.模板匹配的语音识别

在本节中，将构建映射函数$f_1: s\to\hat{y}$，其中 s 是音频信号，$\hat{y}\in\mathbb{R}^n$是单位向量。将y_i表示为 y 的第 i 维，如果$s\in y_i^+$，则y_i= 1;如果$s\in y_i^-$，则y_i= -1;如果$i'\neq i$，$y_{i'}$=0。此外，将u_0定义为表示从语音输入转换的控制值的正值，如果y_i(j)=1(当y_i在时间 j)，则$u_i=u_0$；如果y_i(j)= -1，则$u_i=-u_0$。值得注意的是，在语音输入的时刻只有一个$y_i=\pm1$，$i=1,\ldots,n$。在下一节中，我们将讨论如何获取$\|y\|$，这是$\hat{y}$的度量。

DTW 算法基于动态编程技术。 该算法旨在测量可能在时间或速度上变化的两个时间序列之间的相似性。 在本文中，我们应用 DTW 计算 MFCC 系列的距离向量，用于测量音频信号之间的相似度，以及欧几里得距离来计算音频特征序列的距离向量。 在获得训练模板的当前语音特征序列和每个语音特征序列之间的距离之后，将具有最大相似性的

训练模板中的目标对象分配给测试样本。

语音测量系统

语音中有两种广泛的信息类型。 讲话的语义部分包含客观信息，只要按照语言发音规则进行发音。 另一方面，主观信息指的是隐含的消息，例如说话者的情绪状态或控制意图，这将在本节中进行研究。

A.定量特征提取

可以精确表示控制意图的一组特征对于量化结果的准确性非常重要。两个可能的候选特征如下研究。

1)基于能量的特征

音量表示语音强度，可以表示为每帧中信号的幅度。 这是表达控制意图的最直接的方式。 提取能量特征的基本步骤如下。

步骤 1：框架

将从模数转换获得的语音样本分割成长度在 20 到 40 毫秒范围内的小帧的过程，语音信号被划分为 N 个样本的帧，相邻帧被 M（M < N）。 在本文中，我们选择 M = 128 和 N = 256。

步骤 2：端点检测

语音信号的端点检测是直接影响定量结果准确性的一个步骤。 这里，应用了使用短期能量和短期平均过零率的双阈值语音端点检测算法。

步骤 3：平均活跃能量

两个端点之间的平均能量将被计算为控制意图强度的量化结果，以进一步映射为机器人的控制输入。

这种方法的存在几个缺点：(i）并不是所有人通过改变体量来表达自己的控制意图;（ii）声源必须固定；（iii）它对背景噪音敏感。 然而，优点是（i）不需要模型训练过程；（ii）虽然变化的数量可能不是表达控制意图的自然方式，但在用户知道这一规则的条件下，它是最直接和容易控制的方式。

2）基于频谱的功能

步骤 1：框架（与上述相同）

步骤 2：端点检测（与上述相同）

步骤 3：汉明窗

汉明窗方程给出为$Y(n)=X(n)+W(n)$，其中$Y(n)$是输出信号，$X(n)$是输入信号，$W(n)$是窗口定义为$W(n)=0.54-0.46\cos(2\pi n/(N-1))$，0≤n≤N-1。

步骤 4：快速傅里叶变换

$$Y(w)=FFT\left[H(n)*X(n)\right]=H(w)*X(w) \qquad (1)$$

其中$X(w)$，$H(w)$和$Y(w)$分别是$X(n)$，$H(n)$和$Y(n)$的傅立叶变换。

提取基于频谱的特征的总体进展如图 3 标记为红色部分。与通常使用的光谱特征(例如 LPCC，MFCC 和 LFPC 等）相比，光谱的功率被直接作为后续分类器或回归器的输入，而不是使用一组给定滤波器进行映射，并且采用 DCT 获得功率时域序列。这种改进的动机是，不同的用户可以强调不同的频率来表达他们的控制意图，并且随后的回归者能够学习这种模式。

基于频谱特征的优缺点与能量特征几乎相反。优点：（i）它对背景噪声是鲁棒的；（ii）能够学习人们表达控制意图的自然模式；（iii）声源可以是移动的。缺点：（i）需要培训过程，即应提供大量的培训数据；（ii）训练目标是人为给予的，因此校准误差将是另一个控制误差源；（iii）虽然这种方法可以学习一种自然的控制模式，但是这种模式可能会适应演讲者的情感，环境和时间，很难被刻意表达。

B.语音测量

在这个部分，将要构建地图函数$f_2 : s \to \| y \|$。用来在提取一组可能代表用户控制意图强度的功能之后进行决定，如何将它们映射到控制量级。对于作为标量的能量基特征，仅需要常数系数将所提取的语音特征线性地映射到控制输入中。对于作为特征向量的基于频谱的特征，需要回归步骤。

因此，对于基于频谱的特征，这里应用了一个名为 Random Forest 的回归函数。 它是一种使用决策树作为基本分类器的集体分类。选择RF作为回归因子的几个主要原因：（i）高速度（ii）高精度（iii）评估每个特征变量（iv）能够处理高维特征向量的重要性的能力，这意味着 不需要功能选择。特别是，高计算速度对于保证语音控制系统的实时性能至关重要。

控制系统

本节将介绍几种控制方案，使语音系统的定量结果能够转化为持续的控制输入。

A.系统动力学

考虑线性系统动力学：

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

其中x_1和x_2表示位置坐标，ψ表示运动方向。 将控制输入表示为$u(j) = \left[v(j)\omega(j)\right]^T$。

在本文中，由于环境被假设为机器人未知，机器人无法直接测量控制误差， 但人可以观察到。这意味着机器人只能通过音频传感器的输出来感测环境。 这个假设在现实中可能太不可思议了 然而，它是激励在实际应用中用作辅助校正系统的基于语音的 HRI 系统的开发。

B.控制器设计

实现这一任务的一个常见方法是，每次有演讲指令，机器人转动一个固定的角度，

可以表示为：

$$\omega(j)=c \qquad (3)$$

其中 c 是正常数。 问题是小角度 c 可能导致较慢的控制速度；相反，如果 c 太大，则可能非常难以准确地达到所需的方向。

或者，为了使机器人更好地了解人的命令，在控制错误的条件下，控制规则将被设计用于语音控制系统，只能由人类，而机器人可以通过人类传递的语音信号间接测量控制误差。 表示：

$$f:s\rightarrow y \qquad (4)$$

作为语音系统的融合输出，包括语音识别系统（ $f_1:s\rightarrow \hat{y}$ ）和语音测量系统（ $f_2:s\rightarrow\|y\|$ ）。

将环境噪声和系统噪声综合考虑为ξ，并将$\|y\|$写为$\hat{y}$。语音测量系统的估计进度可以被标记为

$\hat{y}=f_2\left(s\right)+\xi$。 在本节中，噪声被忽略，$\xi$ = 0。进度函数可以重写为$\hat{y}=f_2\left(s\right)$。从观察到的控制误差到控制意图的人脑图功能被定义为$f_3:\psi_e\rightarrow y$，这实际上是未知的。本节的主要目的是对 f3 进行近似，$\hat{y}$是 y 的估计。 我们假设$\hat{y}=y$在这里，这意味着言语系统的结果被假定为等于人的控制意图。

同时，我们假设ψ_e和 y 之间的线性解耦关系，其将时间 j 的映射函数$y=f_3\left(\psi_e\left(j\right)\right)$简化为$y\left(j\right)=\alpha^*\psi_e\left(j\right)$其中$\alpha^*$是未知的正常数。 我们将$\alpha^*$的倒数表示为$g^*$，即$g^*=\alpha^{*-1}$。 类似地，$g^*$是未知的正常数。 那么我们有$\psi_e\left(j\right)=g^*y\left(j\right)$。

人类自适应控制系统。控制规则$\omega\left(j\right)$被选择为：

$$\omega\left(j\right)=k_y y\left(j\right) \qquad (5)$$

其中k_y> 0 是常数。 在这种方法中，给出恒定系数k_y，以将测量结果映射到控制方向误差，那么人脑会经过几次尝试后逐渐找出这个系数。 因此，我们称之为“人为自适应控制”。

C.加速策略

速度$v\left(j\right)$的小值可能有利于方向调整，而较大的$v\left(j\right)$值可能有益于减少任务时间成本。

为了平衡两个因素，在开始时设置相对较小的v_0值，并且引入长度为l的窗函数$W(j)$，

$$W(j)=\begin{cases}1\ if\ 0\le j\le l-1\\0\ \ otherwise\end{cases}\qquad(6)$$

那么加速策略设计为:

$$v(j)=\begin{cases}k_v v(j-1)\ if\ \sum_{i=0}^{j=1}y(j-i)W(i)\le\varsigma\\v_0\ \ otherwise\end{cases}\qquad(7)$$

其中$k_v>1$是加速度的常数集，ς是非常小的正值。 我们不将其设置为零的原因是为了避免噪音的不适。