行政院國家科學委員會專題研究計畫　期中進度報告

虛擬實境之多媒體影音控制技術開發(1/3)

計畫類別：個別型計畫
計畫編號：NSC91-2213-E-009-145-
執行期間：91 年 08 月 01 日至 92 年 07 月 31 日
執行單位：國立交通大學電機與控制工程學系

計畫主持人：林進燈

報告類型：精簡報告
報告附件：出席國際會議研究心得報告及發表論文
處理方式：本計畫可公開查詢

中　華　民　國 92 年 5 月 23 日

虛擬實境之多媒體影音控制技術開發(1/3)

計畫編號：NSC91-2213-E-009-145-

執行期限：91.8.1-92.7.31

主持人：林進燈　國立交通大學　教授

執行機構：國立交通大學電機與控制工程研究所

一、前言

本計畫為「虛擬實境之多媒體影音控制技術開發」，研發重點在於結合各種電機資訊與人工智慧等科技，讓虛擬實境系統具備有智慧型的反應能力，研究的方向有虛擬實境多媒體系統的影像控制、音訊控制與通訊等三項技術的開發與研究。在本期的計畫中，就音訊控制技術開發的部份，包括「不特定語者語音辨識系統核心辨識器」與「環場音效系統之研究」，這裡將針對此兩項主題進行進度報告。關於不特定語者語音辨識系統核心辨識器的部份，主要是研發一技術 --- generalized common-vector (GCV) approach，以解決不特定語者辨識的問題，其重點在於對語音特徵進行特徵分析(eigenanalysis)，並去除語者相關的部份，同時結合決策樹分析以解決前後音相關的問題。關於環場音效系統之研究，主要是探討如何去模擬任一空間中音場之響應並研發相關技術，並將此一技術結合杜比 5.1 聲道系統，創造出虛擬的環場音響聆聽效果。以下將分別針對此二主題進行說明

二、不特定語者語音辨識系統核心辨識器

Abstract

A new speech recognition technique is proposed for continuous speech-independent recognition of spoken Mandarin digits. One popular tool for solving such a problem is the HMM-based one-state algorithm. However, two problems existing in this conventional method prevent it from practical use on our target problem. One is the lack of a proper selection mechanism for robust acoustic models for speaker-independent recognition. The other is the information of intersyllable co-articulatory effect in the acoustic model is contained or not. In this paper, we adopt the principle component analysis (PCA) technique to solve these two problems. At first, a generalized common-vector (GCV) approach is developed based on the eigenanalysis of covariance matrix to extract an invariant feature over different speakers as well as the acoustical environment effects and the phase or temporal difference. The GCV scheme is then integrated into the conventional HMM to form the new GCV-based HMM, called GCVHMM, which is good at speaker- independent recognition. For the second problem, context-dependent model is done in order to account for the co-articulatory effects of neighboring phones. It is important because the co-articulatory effect for continuous speech is significantly stronger than that for isolated utterances. However, there must be numerous context-dependent models generated because of modeling the variations of sounds and pronunciations. Furthermore, if the parameters in those models are all distinct, the total number of model parameters would be very huge. To solve the problems above, the decision tree state tying technique is used to reduce the number of parameter, hence

reduce the computation complexity.

Introduction

Automatic speech recognition (ASR) is useful as a form of input. It is especially useful when someone's hands or eyes are busy. It also allows people with handicaps such as blindness or palsy to use computers. Because of the potential applications mentioned above, we attempt to develop a speaker-independent automatic speech recognition system for Mandarin digits.

In recent years, most automatic speech recognition technologies were based on the so-called Hidden Markov Models (HMM) and used the connected word pattern matching method to achieve continuous speech recognition. There exists many methods to solve the connected word pattern-matching problem. One well-known method is called one-state algorithm. There are two problems in continuous speech recognition based on the one-stage algorithm, one is how to build a reference model to characterize the acoustic feature of speech signal, the other is the information of intersyllable co-articulatory effect in the acoustic model is contained or not.

Due to the first problem mentioned above, the one-state algorithm is sensitive to the reference patterns, and thus the choice of reference patterns is important. One well-known and widely used statistical method of characterizing the spectral properties of the frames of a speech pattern is the HMM approach. The better the HMM models the acoustic signals, the better performance the one-state algorithm can achieve. One of the most important issues of speaker-independent (SI) speech recognition system is the estimation of robust speech model over different speakers. The statistical speech models for each phone unit of the recognition system should be estimated to cover the spectral variations in speech signal caused by intra-speaker differences. In this thesis, we propose a new framework of HMM called *generalized common-vector-based HMM* (GCVHMM) as a reference for speaker- independent automatic speech recognition.

For the second problem, context-dependent model is done in order to account for the co-articulatory effects of neighboring phones. It is important because the co-articulatory effect for continuous speech is significantly stronger than that for isolated utterances. However, there must be numerous context-dependent models generated because of modeling the variations of sounds and pronunciations. Furthermore, if the parameters in those models are all distinct, the total number of model parameters would be very huge. To solve the problems above, the decision tree state tying technique is used to reduce the number of parameter, hence reduce the computation complexity.

GCVHMM

The statistical speech models for each phone unit of the speaker-independent (SI) recognition system should be estimated to cover the spectral variations in speech signals caused by intra-speaker differences. Gülmezoğlu, et al. proposed a common vector approach (CVA) for SI isolated word recognition. In CVA, a common vector that represents common properties of one specific spoken word is obtained by estimating a common subspace. However, CVA needs the impractical assumption that the training data form a set of linearly independent vectors.

In this chapter, we generalize the CVA to relax its constrain and propose a new extension of HMM called *generalized common-vector- based HMM* (GCVHMM). There are two phases in the GCVHMM, extraction of robust features and estimation of HMM. In the first phase, a generalized CVA is developed based on the eigenanalysis of covariance matrix to extract an invariant feature, called generalized common vector (GCV). To relax the linearly independent assumption in the original CVA, we divide the eigenvalues of covariance matrix into two sets such that all the eigenvalues of the first set are greater than those of the second set. The common vector is obtained by projecting feature vectors on the subspace spanned by the eigenvectors whose corresponding eigenvalues are in the second set. In the second phase, the GCVs are used for the estimation of continuous observation density in HMM and form the so-called GCVHMM. In GCVHMM, in addition to the original elements of a traditional HMM, a new element, *GCV transformation matrix*, is added to extract GCV from speech feature vectors. Finally, a re-estimation algorithm based on Baum-Welch method to estimate all the parameters of GCVHMM is derived.

Structure of GCVHMM

In this thesis, a N state, left-to-right continuous observation density HMM, denoted as Ω, is considered. The initial probability for state i is denoted by $\delta_i = P(\theta_0 = i)$, $1 \le i \le N$, and the transition probability from state i to state j by $a_{i,j} = P(\_t = j \mid \_{t-1} = i)$ for $1 \le i, j \le N$. Denote $\phi = \{\phi_i\}_{i=1}^{N}$, and

$A = \{a_{i,j}\}_{i,j=1}^{N}$. For the calculation of the observation density in state i, denoted as $b_i(o_t)$, for observation o_t, the generalized common vector of o_t given the matrix transformation of generalized common vector is first extracted. Then $b_i(o_t) = P(o_t | \bar{}_t = i)$, $1 \le i \le N$ assumed to be a mixture of Gaussians is then given as

$$b_i(o_t) = \sum_{k=1}^{M} c_{i,k} b_{i,k}(o_t), \quad 1 \le i \le N$$

where M is the mixture number, $c_{i,k}$ is the probability of mixture k in state i, and $b_{i,k}(o)$ is the gaussian distribution given by

$$b_{i,k}(o_t) = \frac{1}{\sqrt{(2\pi)^{D_s} |\Lambda_{i,k}|}} e^{-\frac{1}{2}(y_{t,i,k} - \mu_{i,k})\Lambda_{i,k}^{-1}(y_{t,i,k} - \mu_{i,k})}$$

where $D_s = D - D_g$ is the dimension of the extracted GCV $y_{t,i,k}$ from o_t, $y_{t,i,k}$ is the GCV of o_t for mixture k in state i, and $\Lambda_{i,k}$ and $\mu_{i,k}$ are the covariance matrix and mean vector corresponding to mixture k in state i, respectively. $\Lambda_{i,k}$ is assumed to be diagonal, i.e.,

$$\Lambda_{i,k} = \begin{bmatrix} \sigma_{i,k,1} & 0 & \cdots & 0 \\ 0 & \sigma_{i,k,2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_{i,k,D_s} \end{bmatrix}$$

so that $\left|\Lambda_{i,k}^{-1}\right| = \prod_{l=1}^{D_s} \sigma_{i,k,l}^{-1}$. The GCV $y_{t,i,k}$ from o_t for mixture k in state i is defined as

$$y_{t,i,k} = V_{i,k,o_t}$$

where

$$V_{i,k} = \left[v_{i,k,1}, v_{i,k,2}, ..., v_{i,k,D_s} \right]^T$$

is matrix transformation of generalized common vector for mixture k in state i. For convenience in the following derivation, we also define

$$\mu_{i,k} = V_{i,k} \bar{o}_{i,k}$$

then we can write

$$z_{t,i,k} = y_{t,i,k} - \mu_{i,k} = V_{i,k}\left(o_t - \bar{o}_{i,k}\right)$$

Denote $B = \{b_i\}_{i=1}^{N}$ and $\Omega = \{\delta, A, B\}$.

Reestimation algorithm for the parameters of GCVHMM

For an observation sequence $O = (o_1, o_2, \ldots, o_T)$ unobserved state sequence $\Theta = (\bar{}_0, \bar{}_1, \bar{}_2, \ldots, \bar{}_T)$, and unobserved mixture component sequence $K = (k_1, k_2, \ldots, k_T)$, the joint probability density of $P(O, \Theta, K | \Omega)$ is defined as

$$P(O, \Theta, K | \Omega) = \pi_{\bar{}_0} \prod_{t=1}^{T} a_{\bar{}_{t-1}, \bar{}_t} c_{\bar{}_t, k_t} b_{\bar{}_t, k_t}(o_t)$$

where T is the number of observation in O. It follows that the likelihood of O given Ω has the form

$$P(O | \Omega) = \sum_{\Theta} \sum_{K} P(O, \Theta, K | \Omega)$$

where the summations are over all possible state sequences and mixture component sequences.

Given an observation sequence O, the objective is to maximize $P(O|\Omega)$ over all parameters in Ω. It is, however, difficult to solve this problem by directly maximizing $P(O | \Omega)$ over Ω. In this following, we shall use the EM algorithm to estimate the parameters of HMM. The EM algorithm is a two-step iterative procedure. In the first step, called the expectation step (E step), we compute the auxiliary function for the equation

$$Q(\Omega, \Omega') = \sum_{all\,\Theta} \sum_{all\,K} P(O, \Theta, K | \Omega) \log P(O, \Theta, K | \Omega')$$

In the second step, called the maximization step (M step), we find the value of Ω' that maximizes $Q(\Omega, \Omega')$, i.e.,

$$\bar{\Omega} = \arg\max_{\Omega'} Q(\Omega, \Omega')$$

It has been shown that if $Q(\Omega,\Omega') \geq Q(\Omega,\Omega)$, then $P(O|\Omega') \geq P(O|\Omega)$. Therefore, iteratively applying the E and M steps of equations guarantees monotonic increase in the likelihood. The iterations are continued until the increase in the likelihood is less than some predetermined threshold.

From the following decomposition:

$$\log P(O,\Theta,K|\Omega') =$$

$$\log \phi'_{q_0} + \sum_{t=1}^{T} \log a'_{q_{t-1},q_t} + \sum_{t=1}^{T} \log c'_{q_t,k_t} + \sum_{t=1}^{T} \log b'_{q_t,k_t}(o_t)$$

it is straightforward to shown that $Q(\Omega,\Omega')$ can be decomposed into a sum of four auxiliary functions:

$$Q(\Omega,\Omega') = Q_\phi(\Omega,\phi') + \sum_{i=1}^{N} Q_{a_i}[\Omega,\{a'_{i,j}\}_{j=1}^{N}]$$

$$+ \sum_{j=1}^{N} Q_{c_j}[\Omega,\{c'_{j,k}\}_{k=1}^{M}] + \sum_{j=1}^{N}\sum_{k=1}^{M} Q_b(\Omega,b'_{j,k})$$

where

$$Q_\phi(\Omega,\phi') = \sum_{i=1}^{N}\sum_{K} P(O,q_0=i,K|\Omega)\log \phi'_i$$

$$Q_{a_i}[\Omega,\{a'_{i,j}\}_{j=1}^{N}] = \sum_{j=1}^{N}\sum_{t=1}^{T}\sum_{K} P(O,q_{t-1}=i,q_t=j,K|\Omega)\log a'_{i,j}$$

$$Q_{c_j}[\Omega,\{c'_{j,k}\}_{k=1}^{M}] = \sum_{k=1}^{M}\sum_{t=1}^{T} P(O,q_t=j,k_t=k|\Omega)\log c'_{j,k}$$

$$Q_b(\Omega,b'_{j,k}) = \sum_{t=1}^{T} P(O,q_t=j,k_t=k|\Omega)\log b'_{j,k}(o_t)$$

This implies that the four sets of parameters can be independently maximized. The maximization results of first three auxiliary functions are

$$\phi'_i = \frac{P(O,q_0=i|\Omega)}{P(O|\Omega)}$$

$$a_{i,j} = \frac{\sum_{t=1}^{T} P(O,q_{t-1}=i,q_t=j|\Omega)}{\sum_{t=1}^{T} P(O,q_{t-1}=i|\Omega)}$$

$$c_{j,k} = \frac{\sum_{t=1}^{T} P(O,q_t=j,k_t=k|\Omega)}{\sum_{t=1}^{T} P(O,q_t=j|\Omega)}$$

Substituting the following decomposition

$$\log b'_{q_t,k_t}(o_t) = -\frac{D_s}{2}(2\pi) + \frac{1}{2}\log\left(\left|\Lambda^{-1}_{q_t,k_t}\right|\right) - \frac{1}{2} z'^{T}_{t,q_t,k_t} \Lambda^{-1}_{q_t,k_t} z'_{t,q_t,k_t}$$

where

$$z'_{t,q_t,k_t} = y'_{t,q_t,k_t} - \mu'_{q_t,k_t} = V'_{q_t,k_t}(o_t - \kappa'_{q_t,k_t})$$

for $\log b_{q_t,k_t}(o_t)$ and differentiating it with respect to $\kappa'_{j,k}$ and $\Sigma'^{-1}_{j,k,l}$,, we obtain

$$\bar{\kappa}_{j,k} = \frac{\sum_{t=1}^{T} P(O,q_t=j,k_t=k|\Omega)\cdot o_t}{\sum_{t=1}^{T} P(O,q_t=j,k_t=k|\Omega)}$$

$$\bar{\Sigma}_{j,k,l} = \frac{\sum_{t=1}^{T} P(O,q_t=j,k_t=k|\Omega)\cdot z'^{2}_{t,j,k,l}}{\sum_{t=1}^{T} P(O,q_t=j,k_t=k|\Omega)}$$

where $z'_{t,j,k,l}$ is the lth element of $z'_{t,j,k}$.

To obtain the solution for $v'_{j,k,l}$, which is the lth element of $v'_{j,k}$:

$$\left\langle v'_{j,k,l}, v'_{j,k,l} \right\rangle = 1,\ 1 \leq l \leq D_s$$

the constrains

$$\frac{1}{2}\sum_{l=1}^{D_s} \lambda_{i,k,l}\left(\left\langle v'_{j,k,l}, v'_{j,k,l} \right\rangle - 1\right)$$

are added to $Q_b(\Omega,b'_{j,k})$. Then,

$$\frac{\partial Q_b(\Omega, \hat{b}_{j,k})}{\partial \hat{v}_{j,k,l}} = \vec{0}$$

we obtain

$$R_{j,k}\bar{v}_{j,k,l} = \pounds_{j,k,l}\bar{v}_{j,k,l}$$

where

$$R_{j,k} = \sum_{t=1}^{T} P(O_t, j, k_t = k|\Omega)(o_t - \hat{\kappa}_{j,k})(o_t - \hat{\kappa}_{j,k})^T$$

It can be said that $R_{j,k}$ characterizes the variations for mixture k in state j so that it plays the same role as Φ_X in previous section. Thus, the eigenvectors of $R_{j,k}$ corresponding to the eigenvalues of smallest D_s are selected to constitute the GCV matrix transformation for mixture k in state j.

A Hybrid Decision Tree

To overcome this limitation, we have introduced the integrated generalized common vector approach into the conventional HMM in chapter 3, which is better at speaker-independent recognition because of its ability to extract common invariant features over different speakers. Besides modeling acoustic parameters, most of the variations are due to consistent contextual effects in practice. Therefore, we can focus our research on context-based information. Since the co-articulatory effect for continuous speech is significantly stronger than that for isolated utterances, it is important to study the modeling of context-dependent "subword" units. Here, "subword" means "Mandarin digits", which indicate syllable equally.

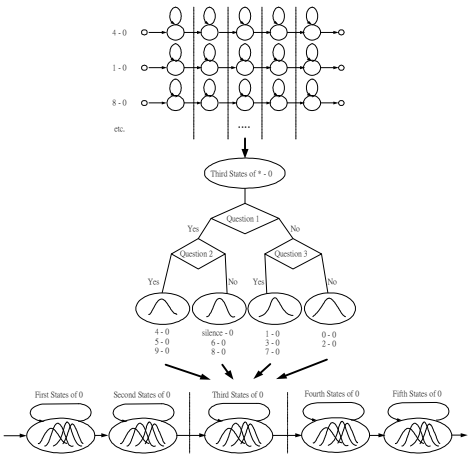
The most important reason why we use the method of decision tree state tying is that the total number parameters in all models is prohibitively large. The computation complexity to train all these parameters would be intolerable. To reduce the total number of model parameters, one approach is to reduce the number of parameters in each model. The way of using continuous HMMs with tied parameters, parameter tying, reduces the parameter count while maintaining the model accuracy, and is popularly used in most ASR systems.

After incorporating common vector features mentioned with the structure of decision tree state tying, the decision tree algorithm should be modified as follows:

1. **Locate a (small) set of left context digit syllable questions manually.**
2. **For each center Mandarin digit syllable p:**
   - Estimate all left context digit syllable GCVHMMs.
   - For each Markov state k in the model topology, classify all the k-th output distribution in all left context digit syllables using a binary tree.
     - **a) Put all the training data in k-th state of all left context digit syllables into the root node.**
     - **b) Classify all the training data by each question in the question set. Using the clustered data to generate Gaussian distribution of common features by the method of GCVHMM introduced in Chapter 3. Then compute the likelihood of the parent node from Equation** (4.15)**:**

       $$L = -\frac{1}{2}\left[D_S\left(1 + \ln(2\pi)\right) + \ln\left(|\Lambda_{i,k}|\right)\right] \cdot \sum_i \sum_t \Upsilon_i(t)$$

       where D_S and $\Lambda_{i,k}$ represent the dimension and the covariance matrix of common vector after the method of GCVHMM.
     - **c) Split the node by each question in the question set. By splitting, some training data that come from the left context digit syllables which answer yes to the question go to the yes-child node; those which answer no to the no-child node. Then calculate the every likelihood of two child nodes. Finally, compute the likelihood increment by each question in the question set.**
     - **d) Find the best question in the question set by computing the most likelihood increment for each of the newly created children.**
     - **e) Go to step b) unless some stop- growing criteria is met.**



GCVHMM of digit syllable "0"

Mandarin Digits Recognition Experiments

The speech data used in our experiments are the set of continuous Mandarin digits. We use a speech database from 20 persons including 10 males and 10 females. Each one speaks 10 times of each Mandarin digit. The recording sampling rate is 8kHz and stored as 16-bit integer.

Balanced Corpora

Digit Model	Decision Tree State Tying	GCVHMMs	HMM
0	91.667	83.333	58.333
1	61.111	27.778	0.000
2	61.538	53.846	76.923
3	100.000	89.474	100.000
4	83.333	50.000	50.000
5	100.000	94.118	100.000
6	60.000	30.000	0.000
7	100.000	46.154	7.692
8	86.667	40.000	60.000
9	100.000	69.231	23.077
Average	84.432	58.393	47.603

Unbalanced Corpora

Digit Model	Decision Tree State Tying	GCVHMMs	HMM
0	75.610	65.854	53.659
1	66.522	76.087	28.261
2	72.727	54.545	54.545
3	84.091	56.818	90.909
4	93.182	100.000	88.636
5	81.818	97.727	95.455
6	75.556	51.111	80.000
7	95.455	100.000	95.455
8	93.182	77.273	90.909
9	90.909	95.455	95.455

Average	82.9052	77.487	76.419

Balanced Tree

Digit Model	Balanced Decision Tree State Tying	Unbalanced Decision Tree State Tying Based
0	75.610	51.220
1	66.522	34.783
2	72.727	63.636
3	84.091	68.182
4	93.182	86.364
5	81.818	77.273
6	75.556	64.444
7	95.455	88.636
8	93.182	77.273
9	90.909	90.909
Average	82.9052	70.272

Conclusion

To consider the contextual effects of continuous speech that play an important role in Mandarin, we combine a method of the Decision Tree State Tying with GCVHMM. The balanced corpora mean that the count of females and males in the database are equivalent entirely. It shows 26.039% improvement when we replace GCVHMM with Decision Tree State Tying based on GCVHMM. Nevertheless, if the database is unbalanced, the performance comparison shows 5.4% improvement by employing the Decision Tree State Tying based on GCVHMM. To overcome leaving the major part of models behind in the unbalanced tree, we modify the tree as the balanced tree. We can find that the results show 12.6332% improvement by employing the balanced decision tree state tying based on GCVHMM.

二、環場音效系統之研究

Abctract

The earliest multichannel reproduction system format was brought up for theatre by Dolby Liboratories Inc in 1950s. The main difference between the conventional stereo (two or three dimension) and multichannel sound system is the setup of surround sound channel. The main purpose of surround channel is to produce the effect of liveness, sense of envelopment, and wide spatiality.

We generate different quality of audio sound sources by a room effect emulator, and then turn the conventional stereo into 5.1 channel sound system by a modified Dolby Surround decoder. We further introduce the concept of room effect impulse response in different room-dimension, and use multi-bands equalizer to generate many kinds of music impression for the multichannel sound system. The total system is a 5.1 channel Multi-band room effect emulator. By the system, we can get a full and a live listening experience.

Introduction

Today, 5.1 channels reproduction system have been frequently used in cinema or home video. One of our purpose is to turn a convention stereo sound into a 5.1 channel sound.

To many people, the term *surround* implies that something new has been added to a stereo [1] audio signal something requiring more than two speakers for reproduction. In 1970’s, Dolby Stereo was established as a stereophonic reproduction system having from three to as many as six channels of sound to enhance the action and drama of theatrical presentations in ways only approached by two-channel systems. The most obvious feature of Dolby Stereo is that an additional channel of sound is reproduced along the sides and back of the theatre to "surround" the audience with sound. Dolby Stereo contains Left and Right channels, with an extra ‘effects’ or ‘surround’ channel [2] in the earliest

incarnations, and more usually, with a centre or ‘dialogue’ channel. These four channels are encoded down to two recorded or transmission channels, and decoded in the cinema for playback. In the mid 1980’s a consumer version of the Dolby Stereo format was developed, called Dolby Surround. One of the main works is to study Dolby surround and generate stereo surround, and then a modified structure, a 2- to- 5.1 channel sound system, is built. [3][4]

In this thesis, we propose a 5.1 channel Multi-band room effect emulator with friendly control interface. In order to get different music quality, we introduce the concept of *room simulation* and *equalizers* into a 5.1 channel sound system.

About room simulation, the impulse response is the result of the many reflections of a sound that occur in a room, and consist in three part, direct signal, early reflections, and fused reflections. The first software implementations of room simulation algorithms were carried out in 1961 by Schroeder. Then an extension of the Schroeder algorithm was by Moorer in 1978. A reverberator introduces a spatial dimension to a piece of recorded sound, which means that it can be used to model a specific acoustic environment in which to affect a dry unaltered signal. Long reverberation times provide the feeling of a large hall, while short reverberation times give the impression of smaller rooms. By a reverberator, we will get a fuller listening experience with spatiality. We referred the structure of Moorer’s reverberator using FIR and IIR filters to make artificial reverberation called a *Reverberator*. An example impulse response for a room is depicted in Fig 1.

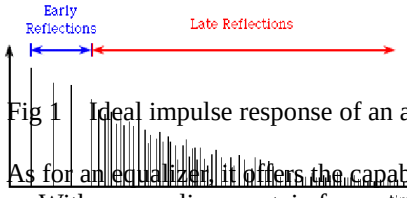


Fig 1 Ideal impulse response of an acoustic room.

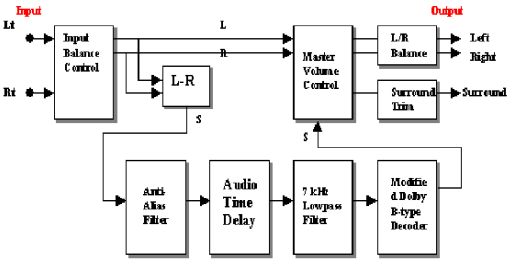
As for an equalizer, it offers the capability of both compensating for defects and fine tuning the system. With an equalizer, certain frequency ranges can be either increased or cut. Referring to a software *Winamp*, we design a 10-band equalizer to control how finely the frequency pattern can be amplified or attenuated, and we setup several selective modes for selection.

At last, we use the concept of fuzzy logic in user interface.

Modified Surround Sound Decoder

The block diagram in Fig 2 shows how the decoder works. The Lt input signal passes unmodified and becomes the left output. The Rt input signal likewise becomes the right output. Lt and Rt also carry the center signal, so it will be heard as a "phantom" image between the left and right speakers, and sounds mixed anywhere across the stereo soundstage will be presented in their proper perspective.

The L-R stage in the decoder will detect the surround signal by taking the difference of Lt and Rt, then passing it through a 7 kHz low-pass filter, a delay line, and complementary Dolby noise reduction. The surround signal will also be reproduced by the left and right speakers, but it will be heard out-of-phase, which will diffuse the image.

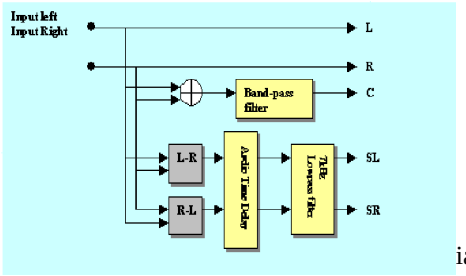


am
oder into a simplified surround sound decoder for
gure 3. As the definition of surround sound described above, this channel is just to present the reverberant effect and feeling of ambiance, but not to present location of sound sources. The terms L-R and R-L referred to Dolby Surround Decoder are to reduce the contents of front channel but not entirely (called leakage). In addition, surround sound sources, L-R and R-L, are out of phase with each other, so the surround channels will diffuse the image in surround sound field. We also use the blocks, *Audio Time Delay* and *7 kHz Low-Pass Filter*, which are described above to make surround sounds more difficult to localize.

the usage of centre channel of cinema reproduction system, the bandwidth of the Band-pass filter is ranging from 20Hz to 20kHz (bandwidth of vocal).

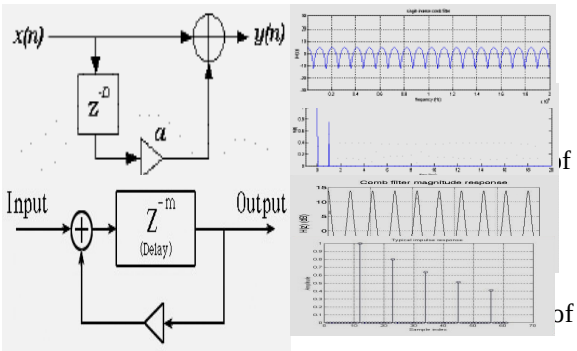
The input left and right siganls are generated from the *Multi-band room effect emulator*, which

will be introduced in Chapter 4. In the next chapter, we shall first describe the basic theory of room effect and basic components of an artificial reverberator.



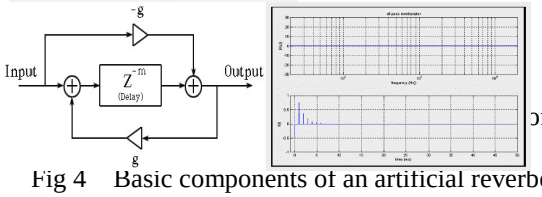
ial Reverberator

The inverse comb filter (FIR), comb filter (IIR) and all-pass filter (IIR) are the basic structures (Fig 4) that have been combined in different ways in an attempt to imitate the effects of various rooms. Fig 5 shows the basic structure of an aitificial reverberator by Schroeder.



f a FIR filter.

of a comb filter.



of a allpass filter.

Fig 4 Basic components of an artificial reverberator.

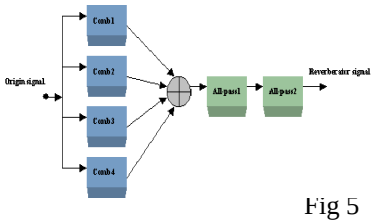
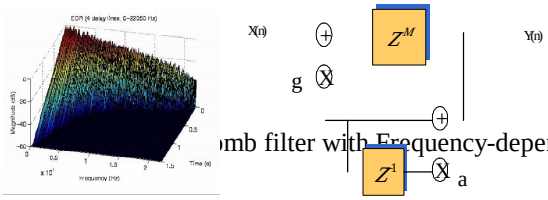


Fig 5 Schroeder's reverberator.

The eigenfrequencies of rooms have a rapid decay for high frequencies. [5][6] A frequency-dependent reverberation time can be implemented with a low-pass filter. Thus Moorer (1978) suggested a modified comb filter with a lowpass filter in feedback loop to take frequency-dependent decay into consideration (Fig 5).



mb filter with Frequency-dependent decay.

According to the concept described above, we use an FIR filter to model early reflections, and an IIR filter that consist of 10 modified comb filters and 4 cascade allpass filters to model the late reflections (Fig 7).

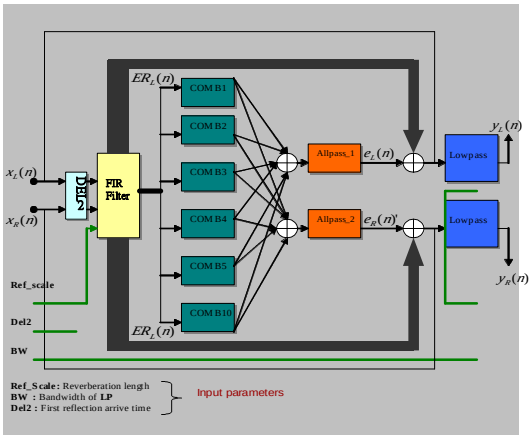


Fig 7 Modified Moorer's reverberator.

In Fig 7, the generated reverberant signals $e_L(n)$ and $e_R(n)$ are added to the direct signals ($x_L(n)$ and $x_R(n)$) and early reflections ($ER_L(n)$ and $ER_R(n)$), and the output signals are:

$$\begin{cases} y_L(n) \leftarrow x_L(n) + ER_L(n) + e_L(n). \\ y_R(n) \leftarrow x_R(n) + ER_R(n) + e_R(n). \end{cases}$$

$\leftarrow$: means that signal path the lowpass filter

The input of the room simulation is the mono signal $x_R(n)$ and $x_L(n)$ respectively. These two mono input signals are added to the left and the right room signals after going through a delay line Del2, and then go through another delay line (*FIR Filter*). The total sum of the early reflections made by *FIR Filter* then goes to parallel circuit of comb filters and cascade allpass filters which implements subsequent reverberation [6].

In order to obtain a high quality spatial impression, it is necessary to decorrelate the room signals $e_L(n) + ER_L(n)$ and $e_R(n) + ER_R(n)$.

10-band Equalizer

Referring to the principles of critical bands [9] and Q equalizer design (1), we design a 10-band equalizer in one octave (Table 1).

Q = f_{centre} / Bandwidth (1)

Band	Frequency (Hz)			Band	Frequency (Hz)		
	Low	High	Width		Low	High	Width
0	0	50	50	5	800	1600	800
1	50	100	50	6	1600	3600	1600
2	100	200	100	7	3600	7200	3600
3	200	400	200	8	7200	14400	7200
4	400	800	400	9	14400	22050	7650

Table 1 The low, high, and width frequencies of each band of the proposed 10-band equalizer

$W_1 \sim W_{10}$ are the input weightings of the 10-bands equalizer, and we can change the weighting of every signal band to generate different kind of music impressions.

The entire structure of the Multi-band rom effect emulator is shown in Fig 8 below.

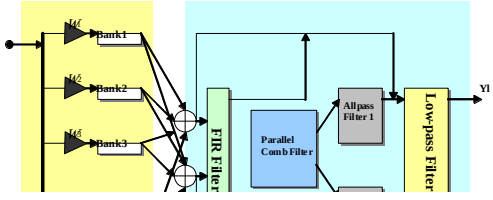


Fig 8 The proposed multi-bands room effect simulator.

Fuzzy User Interface

We introduce the concept of fuzzy logic to generalize the input variables of the proposed *Multi-Bands Room Effect Eimulator*, and build a fuzzy control system as a friendly interface.

the input parameters of the Multi-band room effect simulator can be further divided into two groups. One group is for the decision of room_size, (Fig 9) and the other is for the decision of music impressions. (Fig 10)

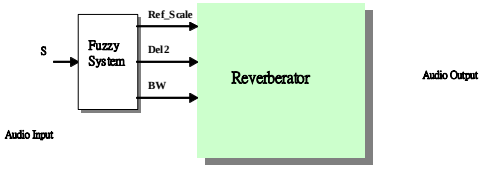


Fig 9 Fuzzification of the input parameters to the reverberator.

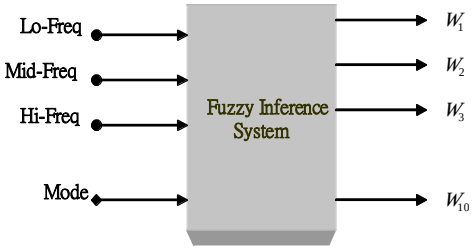


Fig 10 Fuzzy user interface for deciding the band weighting

Experimental Results and Conclusions

All the experiments were carried out in a general listening room. A PC based operation system with a set of 5.1 channel sound system is required. The detail are listed in Table 2 below. Then we let ten classmates grading (1 to 10) the sound effects generated by other room effect generators, and all the users are told the basic theory about room effect at first.

In this thesis, we implement a room effect generator in Visual C++ in a simple way, and then apply it into a 5.1 channel sound system. There are three main problems to be focused on in the future research:

Table 2 Environment requirement.

Software	1. Cool Edit 2000 Syntrillium Software Corporation 2. Winamp Nullsoft, Inc, 3. InterWinRip, 4. Microsoft Windows 2000 professional, 5. Microsoft Visual C++ 6.0
Hardware	1. Pentium IV 1. 5 G 512MB SDRAM 2. Sound Blaster Live 5.1 sound card 3. A pair of 6-pieces loudspeakers 4. eDio AS-100 CineMaster USB Audio box

1. How to reduce the computation loading of FIR filter.
2. No ability to eliminate metallic sound effect.
3. The usage of *Late Low-pass Filter* is improper.
4. How to extract the vocal signal for the center channel.

One of the main problems is about the possibility of real-time processing. This is because that, in our structure, we use a high order FIR filter (more than 2000 orders) to model the early reflection segment, and it would take too much execution time for computing output signal. Lowering FIR order or developing another novel structure will be an important job for real-time realization.

Another problem is about metallic sound effect in long reverberation time. Basically, the *Multi-Bands Room Effect Simulator* won't generate additional metallic sound effect in longer reverberation time. The only case is that, if the original music was heard a little metallic somewhere, the effect still exist in the outputs of the simulator. In other words, the simulator has no ability to eliminate metallic sound effect. For this reason, we can further investigate the way to eliminate metallic sound.

Except to enhance rapid decay for high frequencies, the usage of *Late Low-Pass filter* is to add the feeling of listening position from the stage. The more distant listeners sat, the deeper audio signal they heard. However the usage of enhancing rapid decay for high frequencies is not correct; it suppresses eigenfrequencies of high frequencies unnaturally. Therefore, the *late lowpass filter* is improper for acoustic effect of real situation. Thus we could find another way to model this feature.
At last we can further investigate how to extract vocal signal from music in advance, and then we can use a *Vocal Signal Extraction System* to replace the block, *Band-pass filter*, to generate the center channel of our 2-to-5.1 channel sound system