Voice Activity Detection Using Wavelet-Based Multiresolution Spectrum and Support Vector Machines and Audio Mixing Algorithm

Wei Xue, Sidan Du, Chengzhi Fang, and Yingxian Ye

Department of Electronics Science and Engineering,
Nanjing University,
Nanjing 210093, P.R. China
xwsky2008@hotmail.com,
coff128@nju.edu.cn

Abstract. This paper presents a Voice Activity Detection (VAD) algorithm and efficient speech mixing algorithm for a multimedia conference. The proposed VAD uses MFCC of multiresolution spectrum based on wavelets and two classical audio parameters as audio feature, and prejudges silence by detection of multi-gate zero cross ratio, and classify noise and voice by Support Vector Machines (SVM). New speech mixing algorithm used in Multipoint Control Unit (MCU) of conferences imposes short-time power of each audio stream as mixing weight vector, and is designed for parallel processing in program. Various experiments show, proposed VAD algorithm achieves overall better performance in all SNRs than VAD of G.729b and other VAD, output audio of new speech mixing algorithm has excellent hearing perceptibility, and its computational time delay are small enough to satisfy the needs of real-time transmission, and MCU computation is lower than that based on G.729b VAD.

1 Introduction

Voice activity detection and audio mixing are important techniques in the area of the network speech communication platform. For a wide range of applications such as GSM, Video Meeting, Network Phone, with the use of VAD, system computation load is reduced and network bandwidth is saved. In early VAD algorithms, the VAD determine the silence by comparing the extracted audio features with the initialized threshold. Parameters include short-time zero-crossing, short-time energy, autocorrelation coefficients, LPC [1]-[3]. Because background noise signal is volatile and algorithm uses fixed threshold for comparison, systems adopting these parameters as features perform not well especially in noisy environments.

Papers [4]-[6] suggest using subband or wavelet filters to decompose the audio data, and extract parameters in various frequency. Because of the energy leak of the filter, it is difficult to get the whole frequency spectrum of the signal from the subsection of frequency spectrum. And existing VAD techniques (GSM [7]、G.729B [8], etc.) misjudge part of air stream noise as normal voice. In order to overcome these

T.S. Huang et al. (Eds.): HCI/ECCV 2006, LNCS 3979, pp. 78 – 88, 2006.

problems, we present audio MFCC [9] feature of wavelets-based multiresolution spectrum (WBMS) representing the whole frequency spectrum of the signal, and classify voice by SVM with MFCC of WBMS.

Speaking by more than two users at the same time is allowed in speech communication system, the audio mixing techniques should be introduced. Computer has limited ability to describe the range of sound intensity. Adopting linear superposition as the algorithm of speech mixing in computer, just like what is done in nature, will inevitably cause overflow of superposition [10]. And new noise could be added to output audio by mixing algorithm in [11]. To overcome these defects, the paper presents a new method, adaptive-weighted (AW) algorithm.

The paper is organized as follows. In section 2, we describe communication system framework integrated with voice activity detection and sound mixer. In section 3, we present voice activity detection algorithm including audio feature extraction, silence detection of multi-gate zero cross ratio, SVM theory and VAD algorithm. In section 4, we introduce speech mixing algorithm. Then, section 5 shows experiments on VAD and speech mixing algorithm, and MCU performance. Finally, section 6 gives conclusions.

2 Communication System Integrated with Voice Activity Detection and Sound Mixer

The communication system introduces two new algorithms: voice activity detection of audio sending client, real-time sound mixer of MCU.

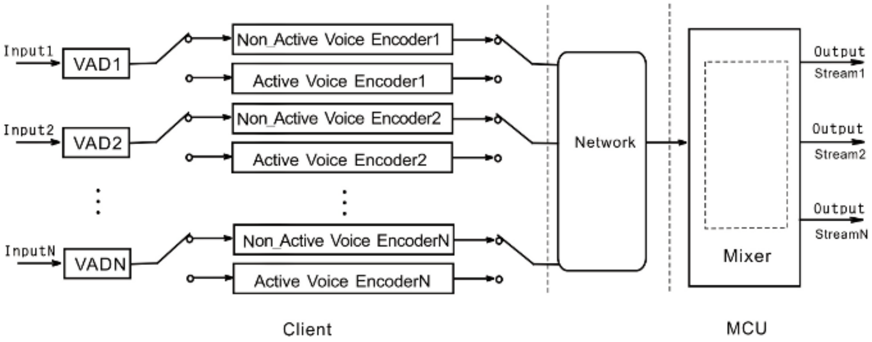

Fig. 1. Proposed communication system framework

The system adopts the centralized communication architecture [12]. According to that, sound mixer processes audio streams in MCU. Audio mixer unit obtains audio streams from every client and sends back to the clients after mixing them. It helps torelease the network communication burden and reduce the computation of every receiving client.

3 Voice Activity Detection

In communication systems based on the agreement such as G.729B, G.723.1A, etc., when the speaker does not speak and there is no obvious noise around, there are still audio frames sent. It shows that the weaker background sound could be judged as the normal voice. For example, the signals of the breath air stream caused by the nose or the mouth near the microphone. So we propose new VAD using detection of multi-gate zero cross ratio and SVM [13] to distinguish the background sound including air stream noise (fake voice) and other background noise from normal voice. In the paper, the sound is divided into three kinds: real silence (background noise), fake voice and normal voice.

3.1 Audio Feature Parameters Extraction

The parameters include two parts: feeling features, MFCC of wavelets-based multiresolution spectrum.

1) Feeling features

Z_{ci}: Z_{ci} is the zero cross ratio of frame i.

E_i : E_i is the short-time energy value of frame i.

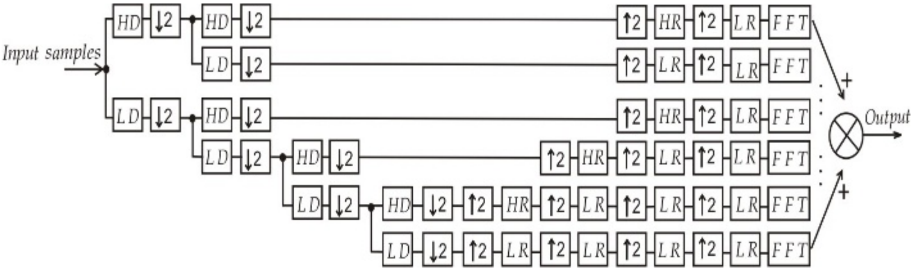

Fig. 2. Direct computation of WBMS

2) MFCC of WBMS

We uses Daubechies4 wavelets [14] to decompose the data into N sub-band, and reconstructs sub-bands to the size the same as that after the first wavelets decomposition, and normalizes each sub-band coefficients, and makes FFT transformation to them. After these, we sum up all the sub-band coefficients altogether to produce WBMS, and then extract MFCC from WBMS.

Suppose LD, HD, LR, HR as the decomposition low-pass filter, decomposition high-pass filter, reconstruction low-pass filter, and reconstruction high-pass filter of wavelets. ↓2 is extraction operator, which extracts the even subscript parts of the original series to produce a new series. ↑2 is filling operator, which inserts a zero in the original series every two points.

MFCC parameters are:

$$c_{MFCC}(i)=\sqrt{\frac{2}{L}}\sum_{l=1}^{L}\log m(l)\cos\left\{(l-\frac{1}{2})\frac{i\pi}{L}\right\} \tag{1}$$

In the equation (1),

$$m(l)=\sum_{k=o(l)}^{h(l)}W_l(k)\left|X_n(k)\right| \qquad ,l=1,2,...,L \tag{2}$$

$$W_l(k)=\begin{cases}\dfrac{k-o(l)}{c(l)-o(l)} & o(l)\le k\le c(l)\\ \dfrac{h(l)-k}{h(l)-c(l)} & c(l)\le k\le h(l)\end{cases} \tag{3}$$

Where o(l), c(l) and h(l) is the lower limit, central limit and upper limit frequency of a triangle filter, X(k) is WBMS.

3.2 Real Silence Detection of Multi-gate Zero Cross Ratio

We introduce multi-gate zero cross ratio [15] to prejudge the real silence. Suppose three different thresholds, $T_1< T_2< T_3$. For every frame, we use equation (4) to compute three multi-gate zero cross ratio Z_1, Z_2, Z_3 with T_1, T_2, T_3.

$$Z_n=\sum\left\{\left|\operatorname{sgn}[x(n)-T_n]-\operatorname{sgn}[x(n-1)-T_n]\right|+\left|\operatorname{sgn}[x(n)+T_n]-\operatorname{sgn}[x(n-1)+T_n]\right|\right\} \\ *w(n-w) \tag{4}$$

The weighted sum denotes the total zero cross ratio.

$$Z= W_1Z_1+ W_2Z_2+ W_3Z_3$$

Where W_1, W_2, W_3 is zero cross weight. Z is the sum of zero cross ratio, which is called weighted sum for short.

By selecting the threshold and weight value properly, we can make the Z of normal voice become obviously larger than that of the silence. Z_0 is defined as the threshold of weighted sum.

When $Z> Z_0$, it is judged as voice frame, otherwise, as silence frame.

3.3 Support Vector Machines

The basic theory of the two classification SVM is, maping the input space to another space (feature space) using non-linear transformation, then seeking the optimal linear classification plane of the samples in this new space (maximizing the classification interval of the two class samples). The non-linear transformation is realized by defining proper inner product function (or kernel function). The two sample classes mentioned above nearest to the optimal classification plane, is called support vector (SV).

Suppose two dividable sample sets (x_i,y_i), i=1,2, …,n, x_i= [x_i^1,x_i^2,…,x_i^d] ^{T}, $x_i \in R^d$, $y_i \in \{+1, -1\}$ is class maker. Optimal classification plane function is:

$$g(x_i)=\sum_{j=1}^{n} a_j^{op} y_j K(x_i,x_j)+b^{op}, \quad i=1,2,\cdots,n \tag{5}$$

Where b^{op} is classification threshold, K(x_i,y_i) is a inner product function. Here we select the radial basis function below as the inner product function. Such SVM is a radial basis function classification machine (in experiment, σ^2=0.3).

$$K(x_i,x_j)=\exp\left(-\frac{\| x_i - x_j \|^2}{\sigma^2}\right) \tag{6}$$

Optimal classification plane function is decided by the optimal results of the function Q(a) below.

$$\underset{a}{Min} \quad Q(a)=-\sum_{i=1}^{n} a_i +0.5\sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j y_i y_j K(x_i,x_j) \tag{7}$$

$$Subject \quad to \quad \sum_{i=1}^{n} y_i a_i = 0, \quad c \ge a_i \ge 0, \quad i=1,2,\cdots,n$$

Where c is a constant. Equation (7) is a quadratic function under inequality constraint, exclusive optimal answer exists. According to the conditions Kühn-Tucker, among the optimal value of Q(a), most a_i^{op} is 0. When the a_i^{op} is unequal to 0 (denotes a_i^{sv}, i=1, 2,…, s), x_i^{sv} is support vector, i=1, 2, …, s, and y_i^{sv} is class marker of x_i^{sv}, b^{op} can be worked out in equation (8).

$$b^{op} = y_i^{sv} - \sum_{j=1}^{n} a_j^{op} y_j K(x_i^{sv},x_j) = y_i^{sv} - \sum_{j=1}^{s} a_j^{sv} y_j^{sv} K(x_i^{sv},x_j^{sv}) \tag{8}$$

We get the optimal classification function of SVM in equation (9).

$$f(x)=sign\{g(x)\}=sign\left(\sum_{j=1}^{n} a_j^{op} y_j K(x,x_j)+b^{op}\right)=sign\left(\sum_{i=1}^{s} a_i^{sv} y_i^{sv} K(x,x_i^{sv})+b^{op}\right) \tag{9}$$

Where sign (.) is a sign function.

3.4 Algorithm of Voice Activity Detection

Algorithm extracts audio feature parameters from voice frame, and detects the real silence by multi-gate zero cross ratio. If Z < Z_0, the frame is judged as real silence, and is filled with self-adaptive noise produced by the CNG of G.729B. Otherwise, the voice data will be classified by SVM into the normal voice and background sound including the fake voice or misjudged real silence. If the background sound is detected, the frame would be filled with self-adaptive noise.

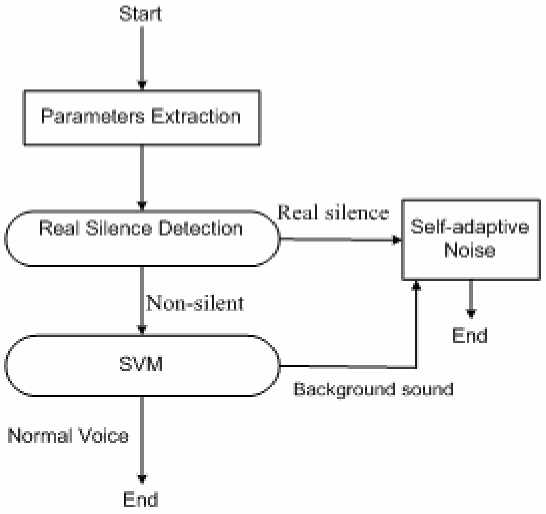

Fig. 3. Voice Activity Detection Algorithm

4 Real-Time Audio Mixing

New algorithm aims at decreasing volume decline of mixed voice data and reducing the spectrum diffused effect of the original signals by slowing down the varying speed of the weight series used by mixer. Define w[j] as weight, which varies only one time during 10 frames, or 0.1s. According to the sampling theory, if effective bandwidth of a series with only 10 samples per second is less than 5Hz, spectrum diffusion can be restricted in the range of ±10 Hz.

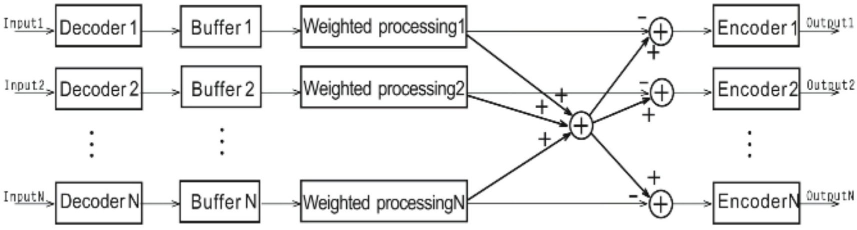

Fig. 4. The principle of speech mixing in MCU

The average value of each audio stream in the ten frames is worked out.

$$Avg[j] = \frac{1}{10l} \sum_{i=0}^{10l-1} \left| data[j,i] \right| \tag{10}$$

Where data[j,i] in the equation (10) denotes the No i sample in the No j audio stream. L is the number of voice samples in a frame. The weight of No j voice stream is computed in equation (11).

$$w[j] = Avg[j] \Bigg/ \sum_{p=0}^{n-1} Avg[j] \tag{11}$$

And then do speech mixing in the equation (12).

$$MixData[i] = \sum_{j=0}^{n-1} data[j,i] * w[j] \tag{12}$$

$$\sum_{j=0}^{n-1} w[j] = 1 \tag{13}$$

Because the computing of Avg[j] of every stream in fig. 4 is independent to each other, each stream can be computed in parallel. This situation continues until reaching the step of speech mixing. So the algorithm is designed into high parallel computing structure, and uses the MMX/SSE/SSE2 instruction set for optimization.

5 Experimental Results

Experiments on VAD compare the proposed VAD with G.729b VAD and MFCC+SVM [4] VAD in terms of probability of detecting speech frames. Experiments on new real-time speech mixing algorithm emphasizes on evaluating the output audio waveform and hearing perceptibility, and how long the speech mixing program consumes when multiple voice streams are being decoded, mixed and encoded in mixer. And MCU performance of video conference system using proposed VAD and new speech mixing algorithm is tested.

Audio signal is sampled at 8kHz and a frame is of 80 samples, each frame continues 10ms. All the data for training and testing SVM of voice activity detection is from database “Aurora”. These sets of data are separately added actually noise or breathing sound. We collected 20 voice samples (60s per sample) for speech mixing test.

5.1 Experiments on New VAD

Multi-gate zero cross ratio for VAD needs optimal weight vectors. We build the target function based on ratio of the incorrect decision to real silence, and traverse each of the weight vector and the threshold value to find out the optimal weight vector and threshold value with lowest ratio of the incorrect decision.

Audio features consist of two parts: two feeling features and L=12 MFCC of WBMS. We adopt SMO [16] algorithm for SVM training. P_d is the probability of correctly detecting normal voice. P_s is the probability of correctly detecting fake noise. P_c is the probability of incorrectly detecting real silence (not including the fake voice). P_f is the probability of correctly detecting silence (including the fake voice and real silence) for the G.729B algorithm.

The results show that the proposed VAD has an overall better performance than MFCC+SVM VAD, G.729B VAD in all SNRs and the noise types used here, as evident by a lower probability of false detection and a higher probability of correct

Table 1. Comparison of the proposed VAD with the MFCC+SVM VAD and G.729B VAD in terms of probability of detecting speech frames for conference noise, MFCC + SVM VAD prejudges real silence by Multi-gate zero cross ratio detection

Noise type	SNR (db)	Proposed VAD			MFCC + SVM VAD			G.729B VAD	
		P_d(%)	P_s(%)	P_f(%)	P_d(%)	P_s(%)	P_f(%)	P_d(%)	P_f(%)
White	25	99.81	95.32	1.48	99.72	88.92	2.36	99.77	16.30
	15	98.47	89.18	2.18	97.43	85.17	3.45	96.61	23.86
	5	94.63	50.28	2.52	91.28	38.48	4.16	85.05	36.75
Conference	25	99.71	93.45	1.72	99.61	87.44	2.52	99.53	19.42
	15	98.26	90.36	2.25	97.27	84.53	3.65	95.76	27.57
	5	93.52	52.63	3.12	90.56	38.23	4.28	84.56	41.83

speech detection. While SNR declining, P_s of proposed VAD has a high fall, but could be above 50%. And high P_f shows G.729B can't efficiently detect fake voice.

5.2 Experiments on New Speech Mixing Algorithm

To analyze the output audio waveform after speech mixing and evaluate hearing feeling, we implement speech mixing algorithm on two audio streams. In fig. 5 (a)-(c), the waveform of mixed result seems the superposition of input streams waveforms, and fig. 5 (d) is similar to (c). In the experiment on hearing perceptibility, the output of AW is superior to that of ASW [17]. Output of AW is smooth and has no plosive and discontinuous sound. Output of ASW has continuous noise, because the weights used in speech mixing varies with the time, and is random, spectrum of output appears too dispersive after speech mixing, and when it comes to 4 mixing streams,

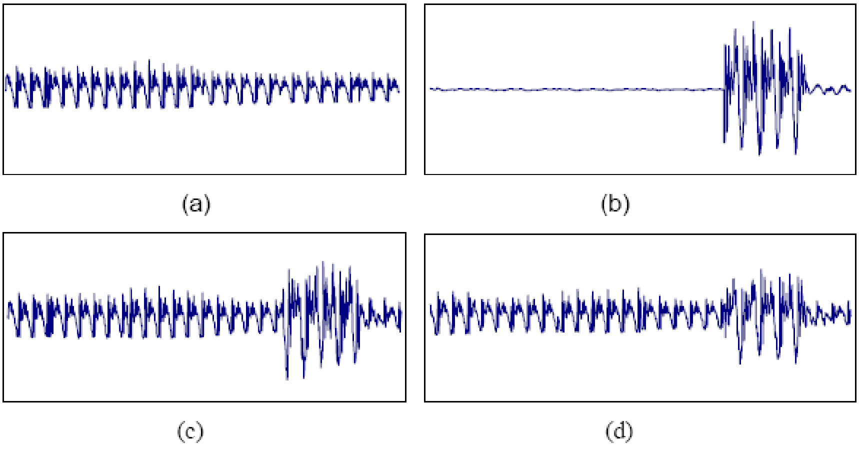

Fig. 5. Mixing result in time-domain ((a), (b) are input audio waveforms in time domain, (c) is output of AW, and (d) is output of ASW)

experiment shows output involves more noise and sounds discontinuous to distinguish the each voice of speaker.

We test how long the mixing program consumes in processing different number streams. In experiments, every stream lasts 60s. Respectively mix the 3-20 audio streams, and decode and encode them by G.729A, results are the mixing time of different number streams. When 20 streams are mixing, the computer's CPU occupation rate is less than 5%.

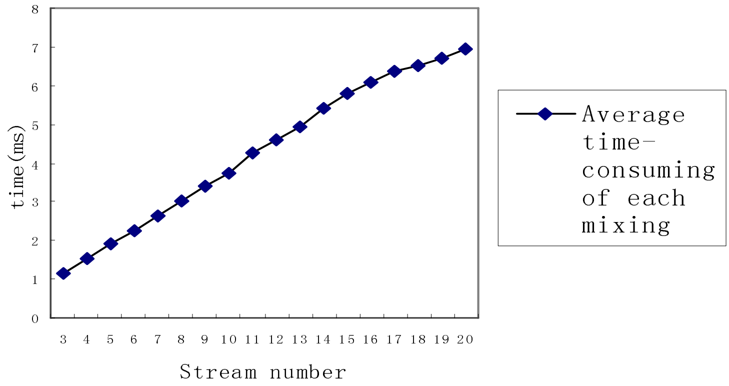

Fig. 6. Time-consuming of different number streams in AW speech mixing (Program runs on a P4 CPU, 512M memory computer)

5.3 Experiments on MCU Performance

We have developed a video conference system using proposed VAD and speech mixing algorithm. For 3-10 conference clients, we respectively calculate the average value of mixed audio streams number of each mixing process of MCU in 20 minutes.

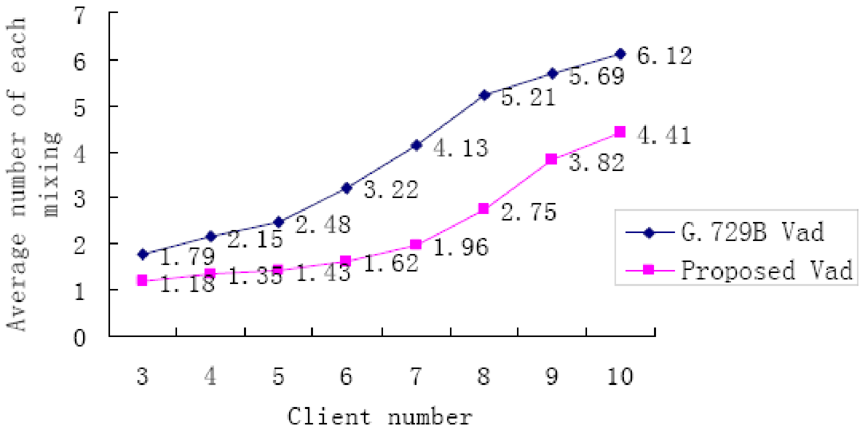

Fig. 7. Comparison of MCU using the proposed VAD with G.729B VAD in average number of mixing streams of each mixing

Fig. 7 shows that the average number of mixing streams of system using proposed VAD, is less than that using G.729B VAD. Because of the high correct decision ratio of proposed VAD, the number of audio data transmitted to MCU decreases, and computation of the MCU declines. System allows more clients to attend the audio discussion in same hardware conditions than that using G.729B.

6 Conclusions

We presents a centralized speech mixing system framework combined with new voice activity detection and new speech mixing algorithm. In the new voice activity detection algorithm, we use MFCC of WBMS and two feeling parameters as the audio features. In order to distinguish the normal voice from the fake voice and real silence, we use the multi-gate zero cross ratio to prejudge the real silence, and complete the voice activity detection by SVM. Compared with the VAD algorithm in the G.729B and the VAD with MFCC+SVM, the proposed VAD algorithm achieves high correct decision ratio of voice in various SNR.

As the important technique of the MCU, AW algorithm is introduced to mix audio streams. Hearing perceptibility test performs well. And the algorithm is designed into high parallel computing structure. For the 20 streams speech mixing, its computational time delay is small and CPU occupation rate is low, the performance of MCU satisfies the needs of real-time transmission. When the video conference system uses the new voice activity detection and AW speech mixing algorithm, speech mixing computation of MCU is much less than that of conference system using G.729B VAD. The same hardware conditions allow more clients to connect MCU and take part in the voice discussion.

References

1. Nemer, E., Goubran, R, Mahmoud, S.: Robust voice activity detection using higher-order statistics in the LPC residual domain. IEEE Transactions on Speech and Audio Processing, vol. 9, March (2001) 217-231
2. J. C. Junqua, B. Reaves, and B. Mak: A study of endpoint detection algorithms in adverse conditions: Incidence on a DTW and HMM recognize. In Proc. Eurospeech'91, (1991) 371-1374
3. A. Sangwan, M. C. Chiranth, H. S. Jamadagni, R. Sah, R. V. Prasad, and V. Gaurav: VAD techniques for real-time speech transmission on the Internet. In IEEE International Conference on High-Speed Networks and Multimedia Communications, (2002) 46-50
4. Guodong Guo, Stan Z. Li: Content-Based Audio Classificationand Retrieval by Support VectorMachines. IEEE Trans. on Neural Networks, vol. 14, no. 1, January (2003) 209-215
5. J. Stegmann, G. Schroeder: Robust Voice Activity Detection Based on the Wavelet Transform. Proc. IEEE Workshop on Speech Coding, September 7-10, (1997) 99-100
6. Chien-Chang Lin, Shi-Huang Chen, T. K. Truong, and Yukon Chang: Audio Classification and Categorization Based on Wavelets and Support Vector Machine. IEEE Transactions on Speech and Audio Processing, vol. 13, no. 5, Sept. (2005) 644-651
7. ETSI: Draft Recommendation prETS 300 724: GSM Enhanced Full Rate (EFR) speech codec, (1996)

8. ITU-T: Draft Recommendation G.729, Annex B: Voice Activity Detection (1996)
9. L. Rabiner and B. H. Juang: Fundamentals of Speech Recognition. Englewood Cliffs, NJ: Prentice-Hall (1993)
10. Agustín JG, Hussein AW: Audio mixing for interactive multimedia communications. JCIS'98, Research Triangle, NC, (1998) 217-220
11. Shutang Yang, Shengsheng Yu, Jingli Zhou: Multipoint communications with speech mixing over IP network. Computer communications, vol. 25, (2002) 46-55
12. Venkat RP, Harrick MV, Srinivas R: Communication architectures and algorithms for media mixing in multimedia conferences. IEEE/ACM Trans. on Networking, vol. 1, no. 1, (1993) 20-30
13. Cortes C, Vapnik C. Support Vector Networks. Machine Learning, vol. 20, (1995) 273-297
14. Daubechies, I., Ten Lectures on Wavelets, SIAM, Philadelphia (1992)
15. Thomas Parsons W.: Voice and Speech Processing. McGraw-Hill Book Company (1986)
16. JOHN C. PLATT: A Fast Algorithm for Training Support Vector Machines. Microsoft Research Technical Report MSR-TR-98-14, April (1998)
17. Fan Xing, Gu Wei-kang: Research on fast real-time adaptive audio mixing in multimedia conference. Journal of Zhejiang University Science, vol 6a, no.6, May (2005) 507-512