VOCAL TRANSLATION FOR MUTENESS PEOPLE USING SPEECH SYNTHESIZER

C.Nijusekar*

P.Jenopaul**

Abstract —

The research perform has enabled a mute man can speak without surgery. An electrode placed on the neck to get the vibration from blabbering voice of the person and also implement the special speech synthesizer for producing him vowels. It possible for the disable person to produce vowels by thinking of them, using a speech synthesizer. In the future, this breakthrough may help erase the word of speech disability.

***keywords*— blabbering voice, speech disability, articulation disorders, speech synthesiser, & average magnitude difference function ,**

* PSN College of engineering & Technology, Tamilnadu, India

** Assistant professor, PSN College of engineering & Technology. Tamilnadu, India

I. INTRODUCTION

The muteness person has inability to speak properly. The Speech of a person was judged to be disordered if the person's speech was not understood by the listener, drew attention to the manner in which he/she spoke than to the meaning, and was aesthetically unpleasant. It also includes those whose speech is not understood due to defects in speech, such as stammering, nasal voice, hoarse voice and discordant voice and articulation defects. Persons with speech disability were categorised as.

- Persons who could not speak at all;
- Persons who could speak only in single words;
- Persons who could speak only unintelligibly;
- Persons who stammered;
- Persons who could speak with abnormal voice; like nasal voice, hoarse voice and discordant voice etc.,
- Persons who had any other speech defects, such as articulation defects, etc.

II MEDICAL TREATMENTS AVAILABLE

For the speech disabled people, it is possible for them to get their voice by surgery- Cochlear Implant surgery. However this surgery can be performed only to speech disabled children below the age of 6. Considering the risk of surgery, it also costs around 8-11 lakhs. So it's practically not possible for every person to get a surgery. So, we engineers have to innovate to help them reach the same level as us in our society.

II PROPOSED SYSTEM FOR PERSONS WHO COULD SPEAK ONLY IN SINGLE WORDS.

Speech sound disorders may be subdivided into two primary types, articulation disorders (also called phonetic disorders) and phonemic disorders (also called phonological disorders). However, some may have a mixed disorder in which both articulation and phonological problems exist. Though speech sound disorders are associated with childhood, some residual errors may persist into adulthood.

A.Articulation Disorders.

Articulation disorders (also called phonetic disorders or simply "artic disorders" for short) are based on difficulty learning to physically produce the intended phonemes. Articulation disorders have to do with the main articulators which are the lips, teeth, alveolar ridge, hard palate, velum, glottis, and the tongue. If the disorder has anything to do with any of these articulators, then it's an articulation disorder. There are usually fewer errors than with a phonemic disorder, and distortions are more likely (though any omissions, additions, and substitutions may also be present)

III PROPOSED SYSTEM

Mutes man can serve as a bridge between two different peoples, the interactive voice response systems connect telephone users with the information they need, from anywhere at any time any language

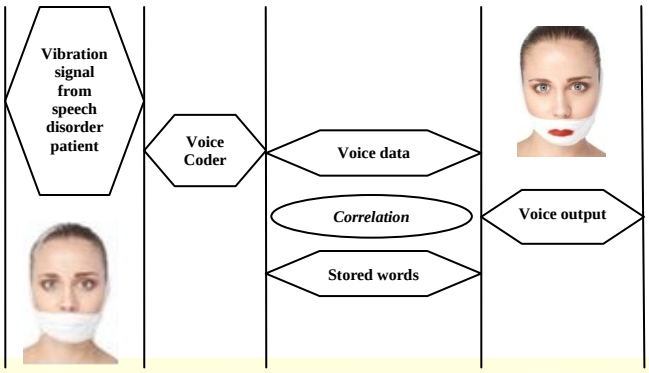

Figure 1. Overview of Proposed System

The project aims to thoroughly explore the theoretical and practical aspects of human vocal translator techniques. To this end a number of specific goals were proposed at the start of the project and detailed knowledge of the challenges faced by speech interactivity embedded modules techniques.

The speech synthesizer is the major role of my project, the Speech synthesis is the artificial production of project is muteness people vocal translation dictionary that is speech synthesizer of disability people speech. A computer system used for this purpose is called a speech synthesizer, and can be implemented in software and hardware.

A. Neural amplifier

Integrated, low-power, low-noise CMOS neural amplifiers have recently grown in importance as large microelectrode arrays have begun to be practical. With an eye to a future where thousands of signals must be transmitted over a limited
bandwidth link or be processed in situ, we are developing low power neural amplifiers with integrated pre-filtering and measurements of the spike signal to facilitate spike-sorting and data reduction prior to transmission to a data-acquisition system. We have fabricated a prototype circuit in a commercially available 1.5µm, 2-metal, 2-poly CMOS process that occupies approximately 91,000 square µm. We report circuit characteristics for a 1.5V power supply, suitable for single cell battery operation. In one specific configuration, the circuit bandpass filters the incoming signal from 22Hz to 6.7kHz while providing a gain of 42.5dB. With an amplifier power consumption of 0.8 µW, the rms input-referred noise is 20.6µV

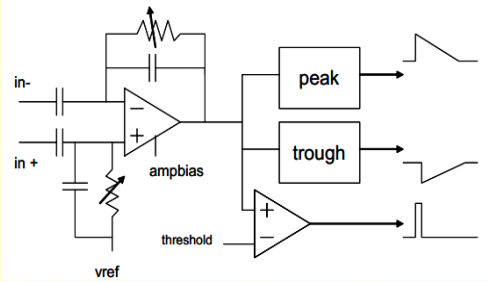

Figure 2. Neural amplifier

1. Frequency Response and Gain

Although we designed the amplifier for a gain of 40dB, we measure a midband voltage gain of approximately 42.5dB (gain = 133). Independent control of the low-frequency corner using rbias and the high-frequency corner using ampbias was confirmed and is shown in Figs. 3 and 4.

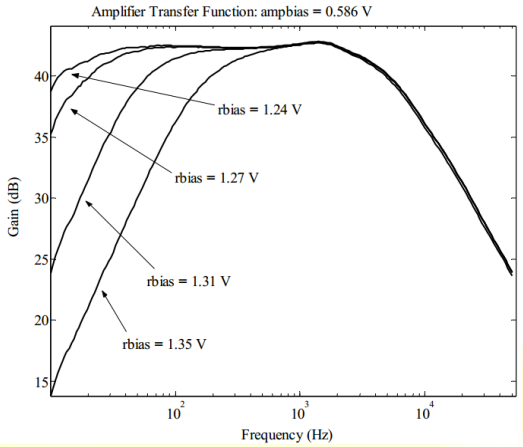

Figure 3. Frequency response of the amplifier as rbias is changed, leaving ampbias fixed at 0.586V. The parameter rbias controls the low-frequency corner of the bandpass filter, making it possible to filter out some of the 60 Hz interference present in any recording situation.

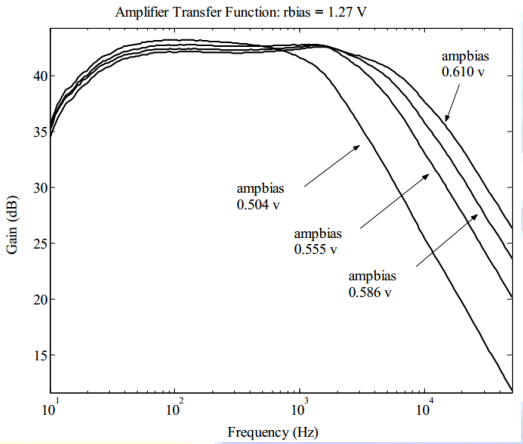

Figure 4. Frequency response of the amplifier as ampbias is changed, leaving rbias fixed at 1.27V. The parameter ampbias controls the high-frequency corner of the bandpass filter.

Using the parameters: ampbias=0.610V and rbias=1.27V, the corner frequencies of the bandpass response (-3dB frequencies) occur at 22Hz and 6.7kHz. Using ampbias = 0.620V & rbias = 1.35V, the corner frequencies occur at 182Hz and 9kHz.

B. Speech synthesiser

Speech synthesis is the artificial production of human speech. A computer system used for this purpose is called a speech synthesizer, and can be implemented in software or hardware products. A text-to-speech (TTS) system converts normal language text into speech; other systems render symbolic linguistic representations like phonetic transcriptions into speech Synthesized speech can be created by concatenating pieces of recorded speech that are stored in a database. Systems differ in the size of the stored speech units; a system that stores phones or diphones provides the largest output range, but may lack clarity. For specific usage domains, the storage of entire words or sentences allows for high-quality output. Alternatively, a synthesizer can incorporate a model of the vocal tract and other human voice characteristics to create a completely "synthetic" voice output.

The quality of a speech synthesizer is judged by its similarity to the human voice and by its ability to be understood. An intelligible text-to-speech program allows people with visual impairments or reading disabilities to listen to written works on a home computer. Many computer operating systems have included speech synthesizers since the early 1990s.

1. Diphone synthesis

Diphone synthesis uses a minimal speech database containing all the diphones (sound-to-sound transitions) occurring in a language. The number of diphones depends on the phonotactics of the language: for example, Spanish

has about 800 diphones, and German about 2500. In diphone synthesis, only one example of each diphone is contained in the speech database. At runtime, the target prosody of a sentence is superimposed on these minimal units by means of digital signal processing techniques such as linear predictive coding, PSOLA or MBROLA. Diphone synthesis suffers from the sonic glitches of concatenate synthesis and the robotic-sounding nature of formant synthesis, and has few of the advantages of either approach other than small size. As such, its use in commercial applications is declining,[citation needed] although it continues to be used in research because there are a number of freely available software implementations.

C. 18-Bit Audio Codec

The PCM3000/3001 is a low-cost, single-chip stereo audio codec (analog – to – digital and digital-to-analog converter) with single-ended analog voltage input and output. Both ADCs and DACs employ delta-sigma modulation with 64-times oversampling. The ADCs include digital decimation filter and the DACs include an8-times oversampling digital interpolation filter. The DACs also include digital attenuation, de-emphasis, infinite zero detection and soft mute to form a complete subsystem. The PCM3000/3001 operates with left-justified, right-justified, I^2or DSP data The PCM3000 can be programmed with a three – wire serial interface for special features and data formats The PCM3001 can be pin-programmed for data formats.

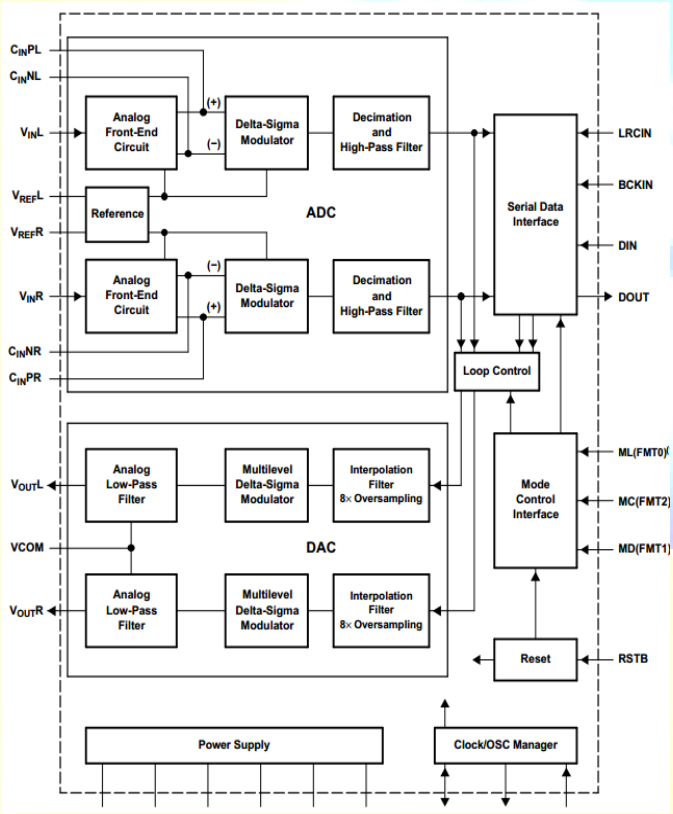

Figure 5.ADC Block Diagram

The PCM3000 ADC consists of a band-gap reference, a stereo single-to-differential converter, a fully differential 5th-order delta-sigma modulator, a decimation filter (including digital high pass), and a serial interface circuit. The block diagram in this data sheet illustrates the architecture of the ADC section.

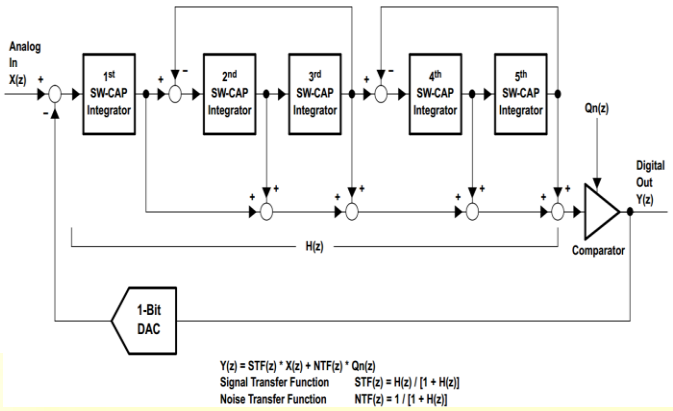

Figure 6. 5th order of ADC Diagram

1. PCM Audio Interface

The four-wire digital audio interface for the PCM3000 DOUT (pin 19). can operate with seven different data for Figure 17. Analog Front-End (Single-Channel)mats. For the PCM3000, these formats are selected through program register 3 in the software mode. For the PCM3000, data formats are selected by pin-strapping the three format pins.

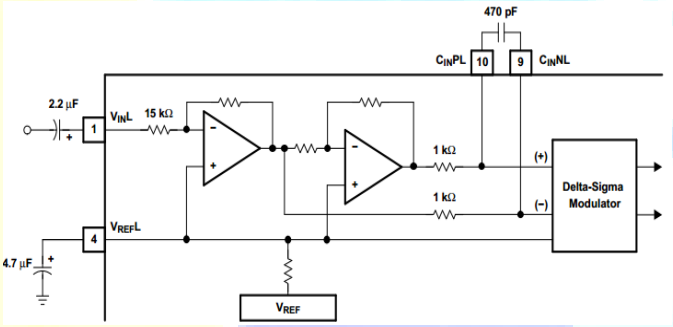

Figure 7. Analog Front-End (Single-Channel)

D. Hardware Architecture

A programmable hardware artefact, or machine, that lacks its software program is impotent; even as a software artefact, or program, is equally impotent unless it can be used to alter the sequential states of a suitable (hardware) machine.

However, a hardware machine and its software program can be designed to perform an almost illimitable number of abstract and physical tasks. Within the computer and software engineering disciplines (and, often, other engineering disciplines, such as communications), then, the terms hardware, software, and system came to distinguish between the hardware that runs a software program, the software, and the hardware device complete with its program.

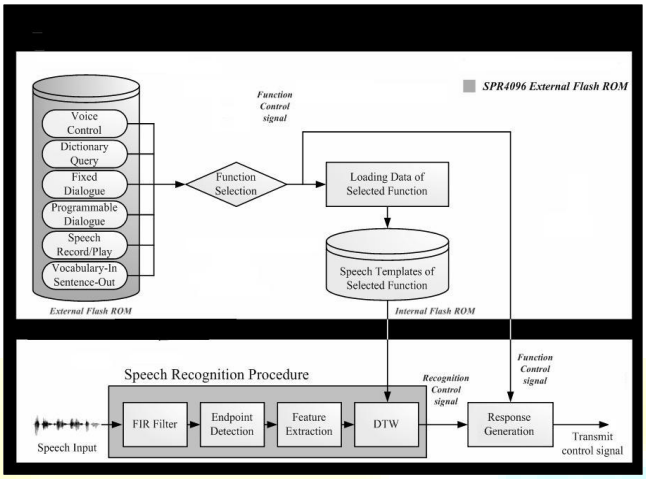

Figure 8. Hardware Architecture of speech synthesizer

1. Voice control

Speaker recognition is the identification of the person who is speaking by characteristics of their voices (voice biometrics), also called voice control There is a difference between speaker recognition (recognizing who is speaking) and speech recognition (recognizing what is being said). These two terms are frequently confused, and "voice recognition" can be used for both.

In addition, there is a difference between the act of authentication (commonly referred to as speaker verification or speaker authentication) and identification. Finally, there is a difference between speaker recognition (recognizing who is speaking) and speaker divarication (recognizing when the same speaker is speaking). Recognizing the speaker can simplify the task of translating speech in systems that have been trained on specific person's voices or it can be used to authenticate or verify the identity of a speaker as part of a security process.

2. Dictionary Query

The Dictionary query is an enquiry form of memory. The memory pre-defined to store all the words to produce the sounds.

3. Fixed Dialog

This query when the user gives the commands it will select the word from the dictionary.

4. Programmable Dialog

If the query is not available the memory it request to him to
Produce the alternate vowels.

E. Simple Manner for Start/End Points Detection

Before processing feature extraction and DTW, we simply detect the start/end points by speech energy (the square value of the acoustic amplitude). Fig. 4 indicates our simple manner to speed up the start/end point detection for recording speech streams with different length. For each speech stream, we arrange equal-size memory spaces for them and initialize their contents as silence data. When a recording process starts, our system start to wait until an energy sum of 80 speech samples exceeds an experimental threshold. And the speech amplitudes sampled from ADC buffer are starting to write into the memory. The end point is detected and the memory writing process is stopped while an energy sum of 80 speech samples is under an experimental threshold.

When the DTW process starts to compute distance between speech stream pairs, we only take the non-silence data from the memory for processing DTW. By using this memory access technique, the start points of different

speech streams can be kept pace for computing DTW distance. This manner can easily record speech streams with different length. Considering to a resource-limited device, it can speed up the start/end point detections for a real-time system.

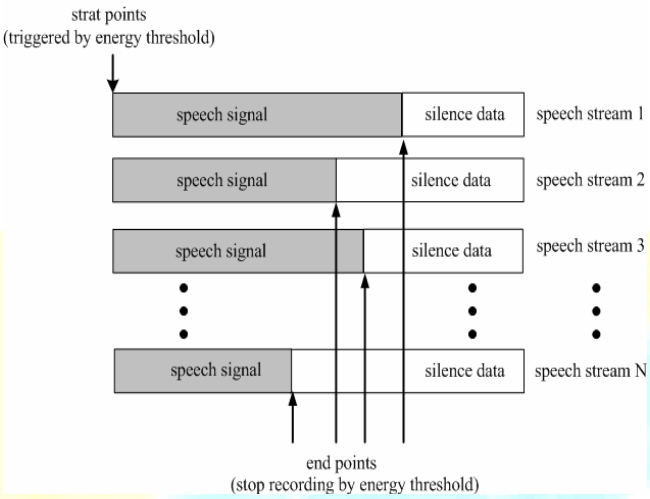

Figure 9. Start/End Points Detection

F. frame-synchronous pitch feature extraction

The original AMDF (Average Magnitude Difference Function) was proposed by Ross et al. in 1974 [6]. It is defined as following

$$P(k)=\frac{1}{N}\sum_{n=0}^{N-1}\left|s_w(\mathrm{mod}(n+k,N))-s_w(n)\right|$$

Where $s(n)$ denotes the speech sample sequence.

In this paper, the circular AMDF (C-AMDF) [7] and modified AMDF (M-AMDF) [8] are integrated to extract circular and modified AMDF (CM-AMDF) pitch features. With the mixed segment containing voiced and unvoiced speech, the CA-MDF can give the pitch period of the voiced part. It is defined as follows:

$$P_C(k)=\frac{1}{N}\sum_{n=0}^{N-1}\left|s_w(\mathrm{mod}(n+k,N))-s_w(n)\right|$$

Where $k = 1,...,N$, N is the length of segment, s (n) w denotes the speech sample sequence with the rectangular window and mod($n + k$,N) represents the modulo operation, meaning that
$n+ k$ modulo N. But sometimes the global minimum valley found from AMDF-based method is not the first local minimum valley especially for voiced speech with good stability and periodicity. To avoid the shortcoming mentioned above, the transform function presented is defined as:

$$P_M(k)=R_{\max}\cdot\frac{N-k}{N-n_{\max}}-P_C(k)$$

Where max {( )} max R P k C = and max n is the index of maximum $PC(k)$. In general, the pitch period is estimated from the short term AMDF as following.

$$T_P=\arg\underset{k}{MIN}^{k_{\max}}_{k_{\min}}(P_M(k))$$

Where max *k* and min *k* are respectively the possible maximum and minimum value. For each pitch of segment *i*, the frequency is calculated by

$$freq(i) = \left\lfloor \frac{S_R}{T_p(i)} \right\rfloor$$

Where *R S* is the sampling rate of speech signal. Fig. 10(b) and (c) show two AMDFs of the same segment. It indicates clearly that the global minimum valley of basic AMDF is 4th local minimum valley, as "P" indicated in Fig. 10(b), whereas the global maximum peak of M-AMDF is first local maximum peak, as "P" indicated in Fig. 10 (c). Since we just search the global maximum peak point of M-AMDF during pitch detection, it can greatly reduce the errors of pitch multiples existed in other AMDF.

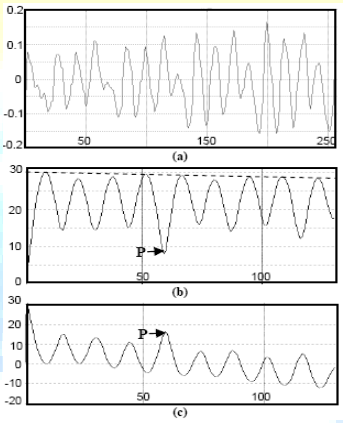

Fig. 10. (a) Original speech, (b) CAMDF for original speech, and (c) MAMDF for CAMDF results.

The flowchart of our frame-synchronous extraction scheduling is shown in Fig. 6. After a speech frame (32 ms) is received from the ADC, both of the CM-AMDF speech feature extraction and memory saving are finished before the incoming of next frame. In our experiments, the total computation time of CM-AMDF computation and memory saving only takes about 0.115 ms which is faster than the speech sampling period (0.125 ms); thus making our system to be workable for real-time demand.

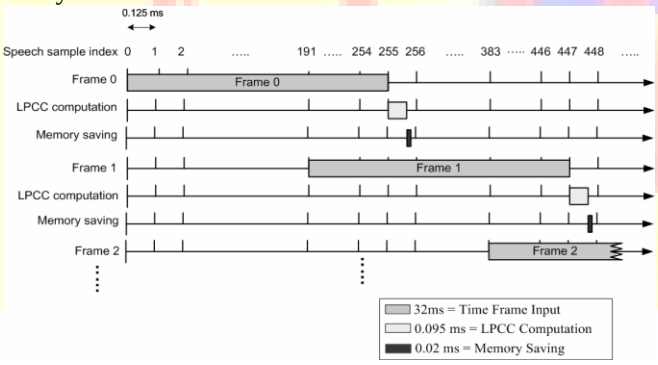

Fig. 11. Program scheduling for real-time speech feature extraction and recording.

G. Dynamic Time Warping Recognizer based CM-AMDF

Template matching has long history in speech recognition application [4]. Basically, it is based on the minimum distance between the test and the reference patterns along the aligned Path, which is obtained using dynamic programming (DP). In the conventional DP-matching technique, the plane of grids shown in Fig. 7 is generally utilized. This figure also shows an example of the

alignment path. Global distance is the distance between the test and the reference patterns along the alignment path, and is computed along the alignment path.

IV CONCLUSIONS

This project can be extended to use voice transmission such as the application of voice recognition for who have specific learning difficulties and also people who do not share a common language. And it can be implement to railway enquiries, airways, conference also applicable for real time communications.

V REFERENCES

[1] C. Kim and K.Seo, "Robust DTW-based recognition algorithm for handheld consumer devices," IEEE Trans. Consumer Electronics, vol. 51 (2), pp. 699-709, 2005.
[2] W. Hess, Pitch Determination of Speech Signal. New York: Springer-Verlag, 1983.
[3] T. E. Tremain, "The Government Standard Linear Predictive Coding Algorithm: LPC-10," Speech Technology Magazine, pp. 40-49, April 1982.
[4] L. R. Rabiner and B. H. Juang, Fundamentals of Speech Recognition.Englewood Cliffs, NJ: Prentice Hall, 1993.
[5] C. Lévy, G. Linarès, and P. Nocera, "Comparison of several acoustic modeling techniques and decoding algorithms for embedded speech recognition systems," Workshop on DSP in Mobile and VehicularSystems, Nagoya, Japan, Apr. 2003.
[6] M. J. Ross et al., "Average Magnitude Difference Function Pitch Extractor," IEEE Trans. Acoust., Speech, Signal Processing, vol. ASSP 22, pp. 353-362, 1974.
[7] Y. M. Zeng, Z. Y. Wu, H. B. Liu, and L. Zou, " Modified AMDF pitchdetection algorithm," IEEE Int. Conf. Machine Learning andCybernetics, pp. 470-473, Phoenix, AZ, Nov. 2003.
[8] Y. M. Zeng, Z. Y. Wu, H. B. Liu, and L. Zou, " Modified AMDF pitchdetection algorithm," IEEE Int. Conf. Acoustics, Speech, and Signal Processing, pp. 341-344, Xi An, China, Nov. 2002.