Real Time Digital Signal Processing

Preliminary Report

Audio Instruction Recognition System for Elevator Application

{ayy17,zy4417}@ic.ac.uk

I. INTRODUCTION

Microphones are widely used to capture sound signals and are deployed in various scenarios to meet the requirement of amplifying and classifying speech. In this project, the aim is to develop a real-time digital signal processing solution for recognising some voice patterns among noises (unwanted voice signals) inside a lift. This process would inevitably involve two topics: 1) Noise cancellation 2) Feature extraction. The first phase would involve random noise detection and reduction. The second phase would require algorithms to classify which floor the lift is (or arriving) at. The following sections would mainly discuss various approaches to tackle the given problem.

II. NOISE SPECTRUM SUBTRACTION

Assuming that additive stationary noise is applied to the audio signal, the sound could be modelled as

$$y(i) = x(i) + N \tag{1}$$

where y is the audio signal detected and N denotes additive noise. However, the noise spectrum remains unknown to the system. Therefore, assumptions are required when adopting this method. This method would work as long as during the first audio frame, silence (no useful information) is recorded. During the first frame, the initial noise spectrum was obtained as S_k where k is the k^{th} frequency component. The output of this process y can be expressed as[1] [2]

$$y_k = (X_k - S_k)\exp^{j\theta(k^{th})} \tag{2}$$

where X represents the frequency domain of the audio signal in convolution with a Hanning window function. Hanning window has a frequency response as shown in Fig 1[3], which mainly works as a low pass filter to suppress high frequency noise.

Another issue that should be taken into account is that a bias is required to suppress negative results from the subtraction in equation 2. This can be simply achieved by assigning "0" to negative results [2].

III. A HYBRID SOLUTION FOR VOICE ACTIVITY DETECTION

A hybrid solution using short-time features of speech frames and a decision strategy to distinguish speech and silence [4] was also discovered. The main idea is to vote on the analysed frames followed by a decision strategy which consists of three different criteria

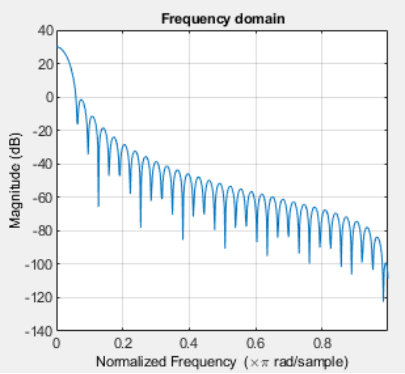

Fig. 1: Frequency response of Hanning window function

– short-term energy threshold, spectral flatness measure threshold, and dominant frequency component threshold. The strategy has been considered appropriate in this application since it is well-established, relatively easy to implement, computationally cheap, and robust in different SNR environments.

The first criterion is short-term energy threshold, which is commonly used in speech/silence detection. This method basically assumes that signals and noise have distinctive features in terms of energy level and therefore they can be separated by analysing the energy spectrum. However, this method is not robust in low SNR environment since the energy of signal and noise are not easily distinguishable, calling for other criteria to be involved.

The second criterion is Spectral Flatness Measure (SFM), which is a measure of the noisiness of spectrum. As its name indicates, SFM reflects the relative flatness of spectrum and can be used to characterize an audio spectrum. For example, a noise, such as white noise, is usually relatively flat in spectrum domain and contain less "spiky values" as the energy is well-spread across different frequency. As shown below, the spectral flatness is calculated by dividing the geometric mean of the power spectrum by the arithmetic mean of the power spectrum[4]:

$$SFM_{db} = 10\log_{10}(G_m/A_m) \tag{3}$$

The last feature is most dominant frequency component of the speech frame spectrum, which can further enhance robustness by distinguishing speech and silence frames. It can be simply computed by finding the frequency corresponding to the maximum value of the spectrum magnitude.

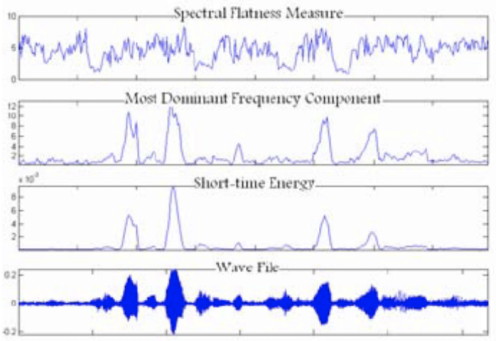

Fig. 2: Speech signal corrupted with babble noise

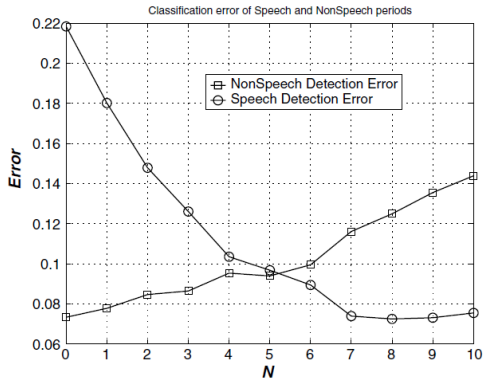

Fig. 3: Error against filter size

Fig 2 [4] provides a good visualisation of each of three features analysing a corrupted signal. As for the actual implementation, an algorithm is proposed but is too long to elaborate in this paper due to page limit. However, it basically follows this procedure: 1) Initialise the threshold for each feature using the first N frames 2) Compute values of each feature for each incoming speech frame 3) Once more than one of the feature values is below the threshold computed previously, this frame would be considered as a speech frame.

IV. Efficient VAD using Long Term Spectral Divergence

Based on the voice activity method mentioned in the previous section, a more efficient approach was founded to increase system performance. This approach has a particular advantage of automatic self-updating in run time, and therefore no reconfiguration is needed for this approach. Long term spectral divergence describes how long term spectral envelope relates to the average noise spectrum. The main advantage of using this method is to highly discriminate various similar audios with noise and to minimize the average number of errors. Long term spectral energy (LTSE) , as shown in equation 4 [5], is used instead of instantaneous spectrum values. Therefore, a moving window with size $2N + 1$ is defined to scan across the frequency domain.

$$LTSE_N(k,l) = \max X(k,l+j)_{j=-N}^{j=+N} \tag{4}$$

where k is the band and l is the frame number currently at. Hence, the N-order long term spectral divergence can be calculated as shown in equation 5 [5],

$$LTSD_N(l) = 10\log_{10}\left(\frac{1}{NFFT}\sum_{k=0}^{NFFT-1}\frac{LTSE^2(k,l)}{N^2(k)}\right) \tag{5}$$

where NFFT denotes the normalized fast Fourier transform and $k \subset \{0, 1....NFFT-1\}$. According to the paper [5] , different orders can be chosen to detect voice activity. Fig 3 [5] indicates that among the different filter sizes, $N = 6$ minimizes the error of detection.

In addition to the mentioned system, another feature can be added to adapt with time-varying noise by continuously

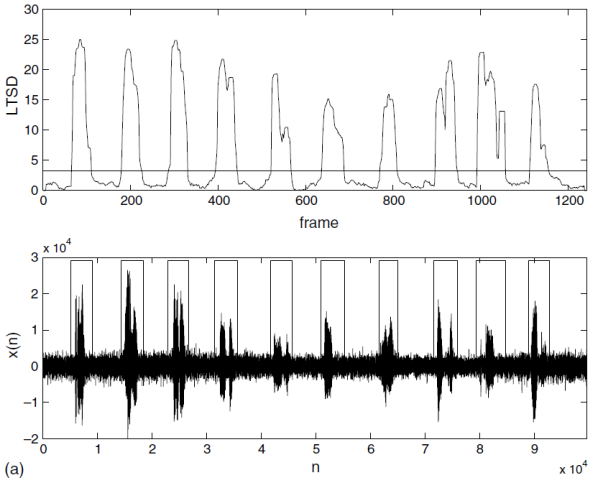

Fig. 4: Effect of LTSD VAD in time domain [5]

updating the noise spectrum N. The following equation [5] denotes how N is related to audio signals.

$$N(k,l) = \begin{cases} \alpha N(k,l-1) + (1-\alpha)N_K(k) \\ \text{if speech pause is detected} \\ N(k,l-1) \\ \text{otherwise} \end{cases} \tag{6}$$

where α is coefficient between 0 and 1 and THRS denotes the threshold for VAD decision.

V. Mel-Frequency Cepstral Coefficient and Dynamic Time Warping

The methods above mainly focus on evaluating if a frame contains voice activity after noise filtering, but they cannot serve to assess voice frame similarity. In this section, discussed is an approaching [6] using Mel-Frequency Cepstral Coefficient (MFCC) for feature extraction and Dynamic Time Warping (DTW) [7] for feature comparisons.

Essentially, recognising targeted signals require a good representation of signals. In our application, the lift is used for humans only and therefore the sounds played are highly likely to be designed for this target group only. In other

words, it has been understood that humans' sensitiveness to different bandwidths is different. Therefore, MFCC would be particularly suitable for feature extraction as its filter bank is designed to mimic human ear perceptions.

Feature Extraction As indicated in Fig 5[6], the feature extraction process involves data pre-processing and Mel Cepstrum analysis.

The main process is: 1) use filter bank to manipulate signals according to human-ear perception 2) convert the spectrum into the log Mel spectrum 3) further convert into time domain using Discrete Cosine Transform and the result is acoustic vector (Mel Frequency Cepstrum Coefficient).

By performing operations above, a good parametric representation of target signal can be obtained and used for later feature matching. Moreover, this solution has authentication capability because MFCC can properly reflect the acoustic characteristics. For example, if there is someone who reads the audio instruction identical to the targeted one, the system will be able to distinguish the "fake" sequences and not to respond.

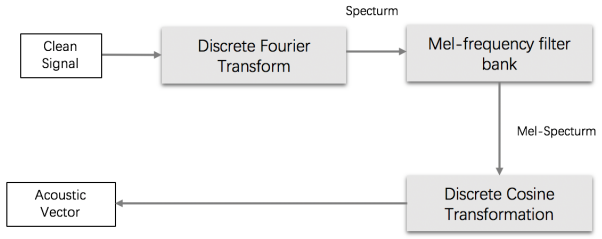

Fig. 5: Feature Extraction Process for Filtered Signal

Feature Matching The feature matching process is based on Dynamic Time Warping (DTW), a method measuring similarity between two sequences which might vary in time or speed. For example, Fig 6[6] demonstrates two signals that are equal in magnitude but are "stretched" along time axis. Based on this method, two sequences containing same information can be effectively compared by calculating the distance between points. Suppose that the target sequence has length n_1 and the input sequence has length n_2, it would be feasible to build a $n_1 \times n_2$ matrix, where each element corresponds to the pairing Euclidean distance as denoted in eq 7 [6] .

$$d(i,j) = (T(i) - I(j))^2 \tag{7}$$

$$D(i,j) = \min[D(i-1,j-1), D(i-1,j), D(i,j-1)] + d(i,j) \tag{8}$$

The minimum distance can be calculated and expressed as shown in eq. 8[6]. In addition, the optimal path can be denoted as:

$$P_o = \arg\min_{i,j} D(i,j) \tag{9}$$

However, this dynamic programming approach suffers from high computational complexity $O(N^2K)$, where N represents the size of input time-domain sequence and K represents the number of ground truth examples stored.

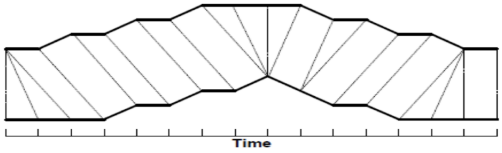

Fig. 6: Two magnitude-identical but time-warped sequences

VI. Overview of system combining all previous methods

After exploring different methods that can serve different purposes, it is important to link them properly and discuss their effectiveness for the desired application. As shown in Fig 7, all previously mentioned methods serve as building blocks for the system tailored to our application.

Applying the basic noise spectrum subtraction in the first step removes all significant noise. Two plausible methods were proposed to detect voice activities in terms of various criterion. The next step would be to suppress time-domain noise by assigning zeros to amplitude of "silent" frames and restore the original phase. At this stage, the gained signal should be noise-free so that MFCC can extract the features of a particular sequence. Finally, dynamic warping with trained audio examples stored would classify if we receive a signal that is similar to target signals and therefore deduces which level was mentioned in the audio.

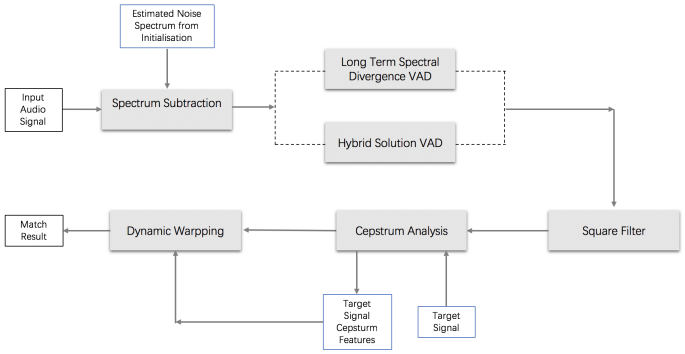

Fig. 7: Flowchart of final design

VII. Conclusion

In theory, the proposed system in this preliminary report should be sufficient in our application of voice pattern recognition among noises. The priorities of algorithms selection and system design have been given to the following criteria: 1) real-time performance (efficient algorithms are chosen) 2) robustness to different noise sources (multiple audio enhancements) 3) authentication between "fake" target signal and "real" target signal. (authentication due to MFCC) 4) no reconfiguration required (chosen algorithms can self-update in run time).

REFERENCES

[1] P. S. Siddala Vihari, A. Sreenivasa Murthy and D. C. Naik, "Comparison of speech enhancement algorithms,"

[2] Steven.F.Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Transactions on Acoustics,Speech and Signal Processing*, vol. 2, 1979.

[3] "Hanning window, https://uk.mathworks.com/help/signal/ref/hann.htm,"

[4] M. H. Moattar and M. M. Homayounpour, "A simple but efficient real-time voice activity detection algorithm," 2009.

[5] C. B. . A. d. l. T. A. R. . Javier Ramırez *, Jose C. Segura 1, "Efficient voice activity detection algorithms using long-term speech information,"

[6] M. B. Lindasalwa Muda and I. Elamvazuthi, "Voice recognition algorithms using mel frequency cepstral coefficient (mfcc) and dynamic time warping (dtw) techniques,"

[7] D. J.Berndt and J. Clifford, "Using dynamic time warping to find patterns in time series," 1994.