# DSP-BASED SPEECH TRAINING OF HEARING IMPAIRED CHILDREN

Dileeka Dias and Ruwan N. Gajaweera
Department of Electronic & Telecommunication Engineering, University of Moratuwa

# DSP-BASED SPEECH TRAINING OF HEARING IMPAIRED CHILDREN

**Dileeka Dias and Ruwan N. Gajaweera**

**Department of Electronic & Telecommunication Engineering, University of Moratuwa**

## ABSTRACT

The paper describes the development of a speech trainer based on digital signal processing (DSP) techniques, for hearing impaired children. The DSP techniques investigated in the implementation of this software-based speech trainer, the final 'product' and its degree of success in actual use are highlighted.

Children with congenital hearing impairments have difficulties in speaking, and even in making the basic sounds associated with speech. Speech therapists use specialized training methods to train such children. The first step in such programs is the training of the pronunciation of common sounds. The dearth of qualified speech therapists, and other facilities hinder the speech development of many children in need of such training.

The computer-based training tool described here, will aid a child, with initial guidance from an adult, to master the pronunciation of initial sounds taught in a speech training programme. An analysis procedure has been developed to compare a hearing-impaired child's utterance with that of a normal person, estimate the degree of correctness and visually indicate this to the child. Through such an iterative visual feedback process, which also indicates the target for correct pronunciation, this software tool can guide the child to self-learn in a game-like environment.

## 1. Introduction

Hearing losses can occur at any age. Some of these conditions are curable and some are not. If hearing loss is present from birth, it is always accompanied by speech loss, unless proper speech therapy is given. There are a few government funded and Non Governmental Organizations (NGOs) helping the training of hearing impaired children. However, such institutions are severely constrained in trained staff and funds to acquire modern training equipment.

**Table 1. Vowel sounds and corresponding symbols**

| Symbol for Vowel Sound | Example |
|---|---|
| ä | father |
| a | man |
| ō | bone |
| ā | fate |
| ü | boot |

During speech therapy sessions, children are first trained to pronounce the long vowel sounds. Table 1 shows the symbols used to represent each of these sounds in this paper.

The speech therapy program then proceeds to consonants, and there onwards to the combination of vowels with consonants. As the next stage two and three syllable words are introduced. Hence, they are gradually taught to communicate with others. Lip-reading helps the trainee to a great extent in these stages. However, the most difficult part comes at the last stage, in the teaching of the manner of articulation. At this stage, lip reading does not help. Therefore, sophisticated training equipment are used today to carry out the entire basic speech training process.

Even though such trainers are extremely expensive, it requires in addition to the software, only a PC with multimedia facilities, which is commonly available at affordable cost.

This was the motivation behind the development of our speech trainer. The software package including the user interface and the underlying DSP techniques were developed to run on common desktop PCs having a sound card and a microphone. This training tool, in its current status, can guide children in pronouncing the five vowel sounds, the first step in a speech therapy course.

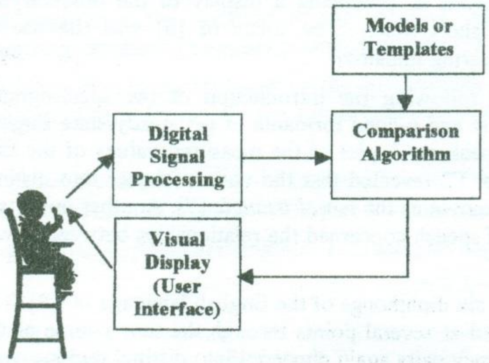


**Figure 1. Basic operation of the speech trainer**

In the training tool developed, the trainee can continuously practice a vowel until he pronounces it correctly. Any deviations from the correct sound are indicated to him visually. Therefore, we first investigated the development of a procedure to compare the trainee's utterance and a template utterance. The user interface, which provides the visual feedback to the trainee was next investigated. Figure 1 illustrates the basic operation of the trainer.

The purpose of the DSP block in Figure 1 is to analyse the trainee's utterance and extract some measures that can be compared with similarly extracted measures from a template. The extracted characteristics should in the first place, be able to differentiate between a correct utterance and one that is not. A secondary, but as important feature of these characteristics should be that they be speaker-independent.

The comparison algorithm compares the characteristics extracted from the trainee's utterance with those extracted in the same way from a normal person's utterance of the same sound. The algorithm examines a set of pre-defined conditions in a pre-determined sequence to arrive at a decision as to whether the trainee's utterance was correct or not.

This decision is conveyed to the trainee through the visual display, accompanied by a visualized version of the characteristics extracted from his utterance. The same characteristics extracted from the template utterance, are also displayed as a target to be aimed at by the trainee.

The major part of the development concentrated on the processing of sounds to extract suitable characteristics and the accompanying comparison algorithms. Different techniques were investigated for this purpose using MATLAB® [1]. The techniques that were found successful were then integrated, using the MATCOM®[2] MATLAB compiler, into a user interface developed through Visual C++ [3].

## 2. Short-Time, Time-Frequency Analysis of Speech

Speech processing techniques for analysis, synthesis, coding, and recognition are well established. It is well known that a speech signal in the time domain, provides little information for parameter extraction. On the other hand, time-frequency analysis has benefited practically every aspect of speech processing through the application of short-time analysis methods [4],[5].

State-of-the art systems for speech recognition, coding, and synthesis segment the speech into short intervals in the order of tens of milliseconds, analysing each segment of speech under the implicit assumption that the signal is stationary over the interval.

Short-time analysis of speech dates back 50 years to the development of *the sound spectrograph* [5]. This provided a means of producing a display of the time-varying spectrum of speech in a relatively short time. The focus of [6] was the use of spectrogram reading as an aid to the hearing impaired.

One of the first large-scale studies following the introduction of the spectrograph, investigated the frequencies of the first and second formants of ten steady-state English vowels recorded from a number of speakers. A plot of the measured values of the first formant F1 versus the second formant F2 revealed that the vowels cluster into distinct regions, defining what is commonly known as the *vowel triangle* [7]. Another important study of the spectral characteristics of speech concerned the relationships between vowel formants and bandwidth [8].

Spectrograms were also used to study six diphthongs of the English language in [9]. The trajectories of F1 and F2 were sampled at several points through the time course of the diphthongs. Scatter plots of the frequency pairs again clustered into distinct regions, with the time course of formant transition in the F1-F2 plane providing further information for analysis.

The notion of time-varying or instantaneous frequency were implicit in all these analyses.

## 3. The Spectrogram

The general trend in speech recognition has been to use formant analysis for the identification of steady-state vowels, and to use spectrographic analysis for consonant-vowel (CV) transitions.

However, during the course of our investigations, it was observed that the speech characteristics obtained through the above two methods from hearing impaired people, differed considerably from those of normal people. This lead to difficulties in using formant analysis and the corresponding vowel triangle as a training tool. Better results were evident through processing of the Spectrogram. Therefore, the Spectrogram was used as the basic tool for extracting sound characteristics for the speech trainer.

The Spectrogram is a method of displaying results of time-dependent frequency analysis of signals. They are typically used in the analysis of non-stationary signals such as Radar, sonar and speech [10].

The time-dependent Fourier transform of a signal *x[n]* is defined as:

$$X[n,\lambda) = \sum_{m=-\infty}^{\infty} x[n+m]w[m]e^{-j\lambda m} \qquad (1)$$

Where *w[n]* is a window sequence. In this analysis, the one-dimensional sequence *x[n]* is converted into a two-dimensional function of the time variable *n*, which is discrete, and the frequency variable $\lambda$, which is continuous.

The above equation can be interpreted as the Fourier transform of *x[n+m]* as viewed through the window *w[m]*. The process can be visualized as the signal passing through a stationary window, so that for each value of *n*, a different portion of the signal is viewed.

The primary purpose of the window is to limit the extent of the sequence to be transformed for each *n*, so that the spectral characteristics of the signal are resonably stationary over the duration of the window. As the window becomes shorter, the frequency resolution becomes poorer, while the time resolution improves.

This time-frequency trade-off of the spectrogram is a well-known factor in short-time analysis. The more rapidly the signal characteristics change, shorter the window should be.

In digital signal processing the Discrete Fourier Transform (DFT) is used extensively, for the computation of which, a variety of different algorithms are available.

Suppose the window has length *L*, with samples beginning at *m=0*, i.e.;

$$w[m] = 0 \quad \text{outside} \quad 0 \le m \le L-1$$

If we sample the $X[n,\lambda)$ at *N* equally spaced frequencies $\lambda_k = 2\pi k/N$, with $N \ge L$, we obtain the discrete, time-dependent Fourier transform from (1) as:

$$X[n,k] = \sum_{m=0}^{L-1} x(n+m)w(m)e^{-j(2\pi/N)km}, \qquad 0 \le k \le N-1 \tag{2}$$

This is the DFT of the windowed sequence *x[n+m] w[m]*. It can also be expressed as:

$$X_r[k] = X[rR,k] = X[rR, 2\pi k/N] = \sum_{m=0}^{L-1} x(rR+m)w(m)e^{-j(2\pi/N)km[k]} \tag{3}$$

Where *r* and *k* are integers such that, $-\propto < r < \propto$ and $0 < k < N-1$.

The above notation denotes explicitly that the sampled time-dependent Fourier transform is simply a sequence of N-point DFTs of the windowed signal segments:

$$x_r[m] = x[rR+m]w[m] \qquad -\propto < r < \propto \text{ and } 0 < k < N\text{-}1. \tag{4}$$

Equation (3) involves the following integer parameters: window length *L*, number of samples in the frequency dimension *N*, and the sampling interval in the time dimension *R*.

The choice $L \le N$ guarantees that we can reconstruct the windowed segments $x_r[m]$ from the block transforms $X_r[k]$. If $R<L$, the segments overlap, but if $R > L$, some samples of the signal are left out, and therefore, it cannot be reconstructed. In general, the three sampling parameters should satisfy the relation $N \ge L \ge R$.

Computational time and resources must also be taken into consideration in determining these parameters, as the spectrogram would be an array of *N* rows and a number of columns determined by the length of the speech segment being analysed.

Figure 2(a) depicts the lines in the $(n,\lambda)$-plane corresponding to $X[n,\lambda)$. Figure 2(b) depicts the sampling points corresponding to $X_r[k]$ in the $(n,\lambda)$-plane for the case $N = 10$ and $R = 3$.

Figure 3(a) shows a spectrogram of a speech segment with $L = 108$ and $R = 16$. This is called a *wideband spectrogram*, as the relatively small *L* implies low time resolution. Figure 3(b) shows the spectrogram of the same speech segment with $L = 720$ and $R = 16$. The larger *L* results in a higher spectral resolution in this *narrowband spectrogram*.

## 4. Spectrographic Speech Processing for the Speech Trainer

Spectrogram of the Sound

The Spectrogram described in the previous section was used as the first step in the digital signal processing for extraction of vowel sound characteristics, which could represent a measure of the correctness of the sound, yet ignore speaker-dependent variations.

Figure 4 illustrates the procedure for the spectrogram computation. The following parameters were used to obtain the spectrogram:

| | |
|---|---|
| Number of samples for FFT (*N*) : | 512 |
| Time sampling (*R*): | 256 |
| Window length (*L*): | 512 |
| Window type: | Hanning |

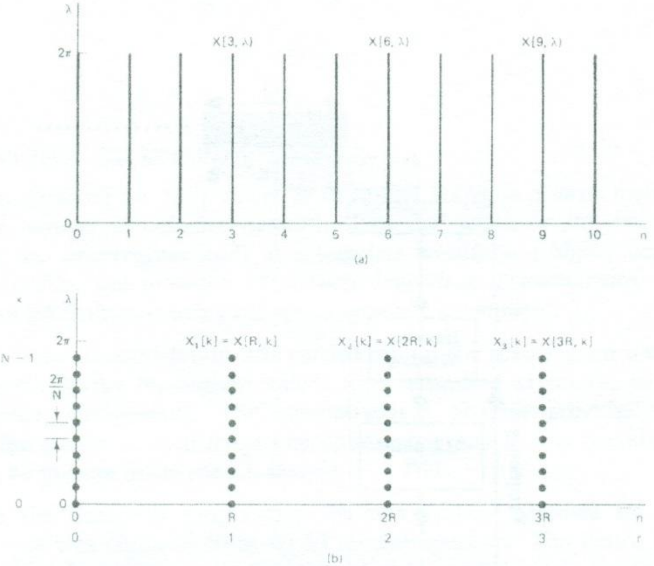


**Figure 2. Grid of sampling points in the *(n,λ)* plane for the sampled time-dependent Fourier Transform with *N = 10* and *R = 3***

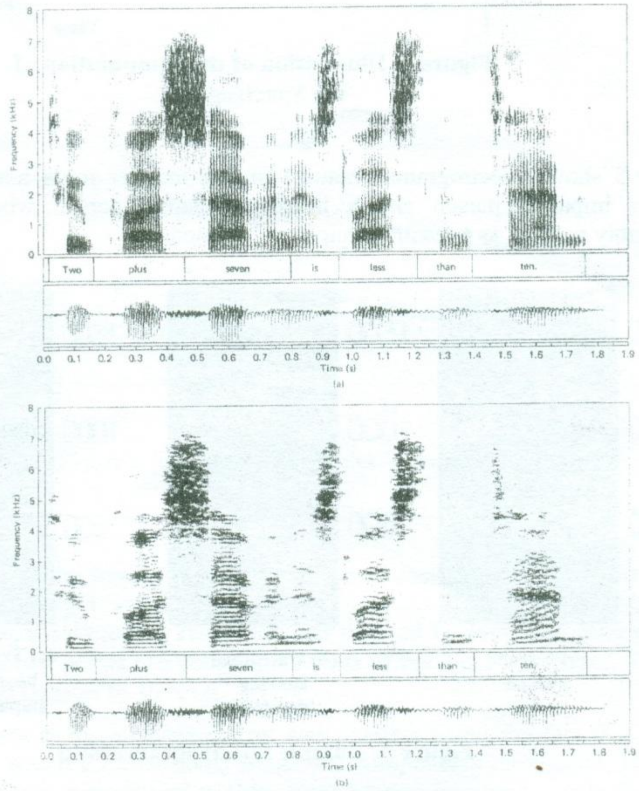


**Figure 3. (a) Wideband spectrogram (b) Narrowband spectrogram for a speech segment**

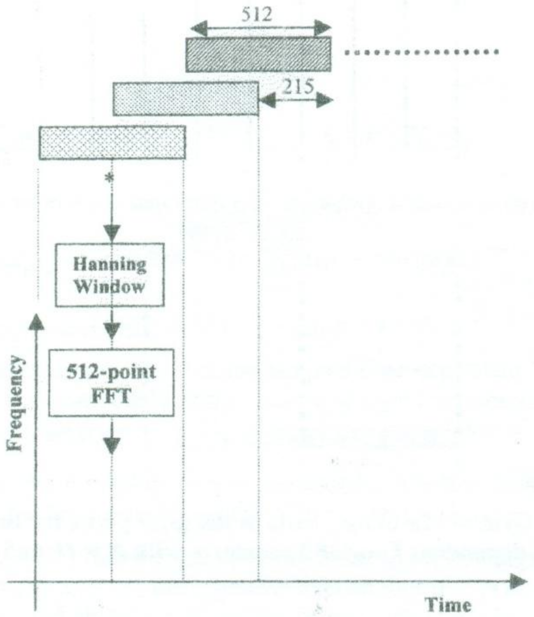
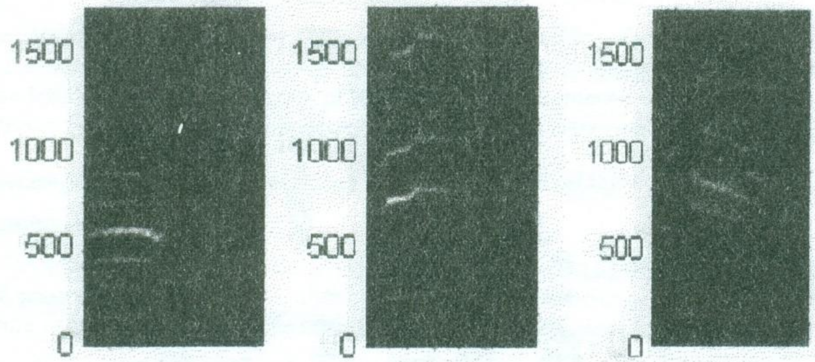


**Figure 4. Illustration of the computation of the Spectrogram**

Figure 5 shows spectrograms obtained in this manner for a normal person, a hearing impaired person and a hearing impaired person whose speech is reasonably normal, as a result of being well-trained.



(a) Normal Person

(b) Untrained hearing impaired

(c) Trained hearing impaired

**Figure 5. Sample Spectrograms for *ä***

Extraction of Template Data from the Spectrogram

The Spectrogram obtained for each vowel as described above is a large matrix, having 256 rows, and a number of columns determined by the length of the speech sample. Therefore, using the spectrogram itself as a template would be a highly computation-intensive task. Further, the presence of speaker-dependent characteristics, and silent periods also cause difficulties in using the spectrogram as a template.

Therefore, a row-wise autocorrelation was carried out on the spectrogram matrix, and a few of the elements having the highest values were identified as corresponding to the dominant frequency components. The autocorrelation process provides a quantity proportional to the power at each frequency component, and it also cancels out silent periods that may be present in the speech sample.

Figure 6 shows the dominant frequencies for two vowels obtained by the above procedure. The data was obtained from 40-50 normal speakers. The figure shows that though the dominant frequency varies from speaker to speaker, they lie in noticeably different frequency bands. The dominant frequency for *ä* falls in the 600 – 1000 Hz band, while that of *ü* falls in the 350 - 600 Hz band. However, there is some overlap due to speaker dependancies.

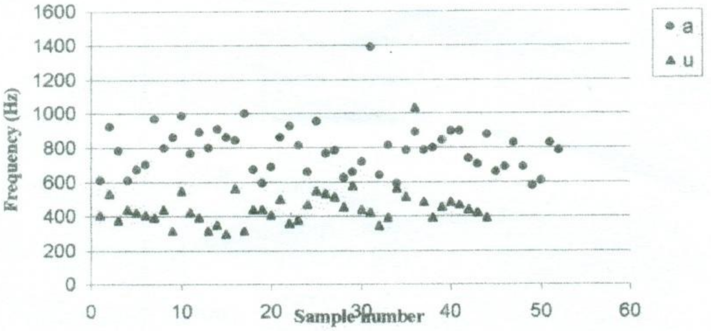


**Figure 6. Dominant frequencies for vowel sounds**

Initial observation of the five vowel sounds indicated that the dominant frequencies fall into different bands, and it would be sufficient to consider 16 frequency bands in the 0 – 4 kHz range. Thus, the frequency resolution was reduced by averaging each group of 16 consecutive frequency points (16 consecutive rows of the analysis matrix). The analysis matrix at this point reduces to 16 rows.

A further reduction in the analysis matrix was possible using the fact that a correctly articulated vowel sound contains fixed frequency components throughout, as illustrated in Figure 5(a). A vast majority of samples from hearing impaired people demonstrated time-varying frequency components. Therefore, the presence of constant significant frequency components throughout the vowel duration was taken as an indication of correct pronunciation. Therefore, the average value of each row of the 16-row analysis matrix was computed, reducing the analysis matrix to a 16 x 1 column vector. The analysis vector (template) for each vowel was obtained by averaging the above column vector for 50 normal, male speakers.

Therefore, the template for each vowel consists of quantities proportional to the power in each of the16 frequency bands when pronounced properly.

Table 2 shows the analysis vectors so obtained, along with computations of the percentage of power in each frequency band for each vowel.

Table 2. Analysis vector for each vowel (power in each frequency band) and the percentage

| Freq. Band (Hz) | ä | % | a | % | ō | % | ü | % | ā | % |
|---|---|---|---|---|---|---|---|---|---|---|
| 0 - 250 | 398.37 | 22.70 | 368.64 | 17.55 | 143.09 | 13.13 | 173.25 | 23.88 | 259.91 | 22.67 |
| 250 - 500 | 65.97 | 3.76 | 138.18 | 6.58 | 503.82 | 46.24 | 461.52 | 63.60 | 603.54 | 52.64 |
| 500 - 750 | 392.20 | 22.34 | 354.47 | 26.40 | 307.34 | 28.21 | 61.24 | 8.44 | 129.51 | 11.30 |
| 750 - 1000 | 506.93 | 28.88 | 392.48 | 18.69 | 106.89 | 9.81 | 23.63 | 3.26 | 5.38 | 0.47 |
| 1000 -1250 | 269.29 | 15.34 | 84.07 | 4.00 | 19.38 | 1.78 | 2.25 | 0.31 | 2.19 | 0.19 |
| 1250 -1500 | 70.05 | 3.99 | 52.27 | 2.49 | 2.00 | 0.18 | 1.44 | 0.20 | 3.41 | 0.30 |
| 1500 -1750 | 7.88 | 0.45 | 141.87 | 6.75 | 1.55 | 0.14 | 0.69 | 0.10 | 5.5 | 0.48 |
| 1750 -2000 | 4.61 | 0.26 | 154.80 | 7.37 | 1.10 | 0.10 | 0.32 | 0.04 | 19.78 | 1.73 |
| 2000 -2250 | 5.42 | 0.31 | 64.92 | 3.09 | 0.84 | 0.08 | 0.23 | 0.03 | 35.88 | 3.13 |
| 2250 -2500 | 4.04 | 0.23 | 48.30 | 2.30 | 0.55 | 0.05 | 0.21 | 0.03 | 31.31 | 2.73 |
| 2500-2750 | 5.23 | 0.30 | 38.39 | 1.83 | 0.60 | 0.06 | 0.24 | 0.03 | 21.92 | 1.91 |
| 2750-3000 | 9.10 | 0.52 | 19.47 | 0.93 | 0.95 | 0.09 | 0.31 | 0.04 | 9.59 | 0.84 |
| 3000-3250 | 9.91 | 0.56 | 25.35 | 1.21 | 0.93 | 0.09 | 0.1 | 0.01 | 8.44 | 0.74 |
| 3250 -3500 | 3.96 | 0.23 | 11.79 | 0.56 | 0.30 | 0.03 | 0.09 | 0.01 | 5.77 | 0.50 |
| 3500 -3750 | 1.70 | 0.10 | 3.76 | 0.18 | 0.17 | 0.02 | 0.07 | 0.01 | 3.13 | 0.27 |
| 3750 -4000 | 0.58 | 0.03 | 1.49 | 0.07 | 0.05 | 0.00 | 0.04 | 0.01 | 1.29 | 0.11 |

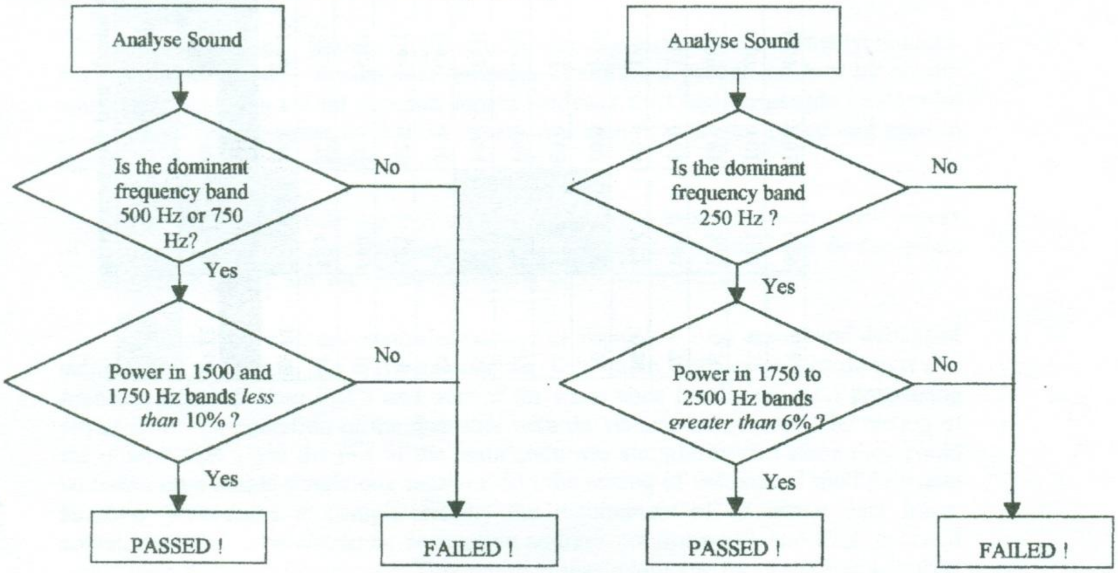
The following features can be extracted fromTable 2:

1. *ä* has most of its power concentrated in the 750 Hz band, with the 500 Hz band closely following.
2. *a* has most of its power concentrated in the 500 Hz band,
3. *a* has more than 10% of its total power concentrated in the 1500- 2000 Hz band, which is not the case for *ä* .
4. *ō*, *ü* and *ā* all have their dominant frequencies concentrated in the 250 Hz band.
5. *ā* has more than 8% of power concentrated in the 1750 – 2500 Hz bands, which is not the case for *ō* and *ü*.
6. *ō* has more than 20% of its power concentrated in the 500 Hz band, which is not the case for *ā* and *ü*.

Factors 1, 2 and 3 above may be used to differentiate between *ä* and *a*. Factors 4, 5, and 6 may be used to differentiate between *ō*, *ü* and *ā*

### 5. Comparison Algorithms

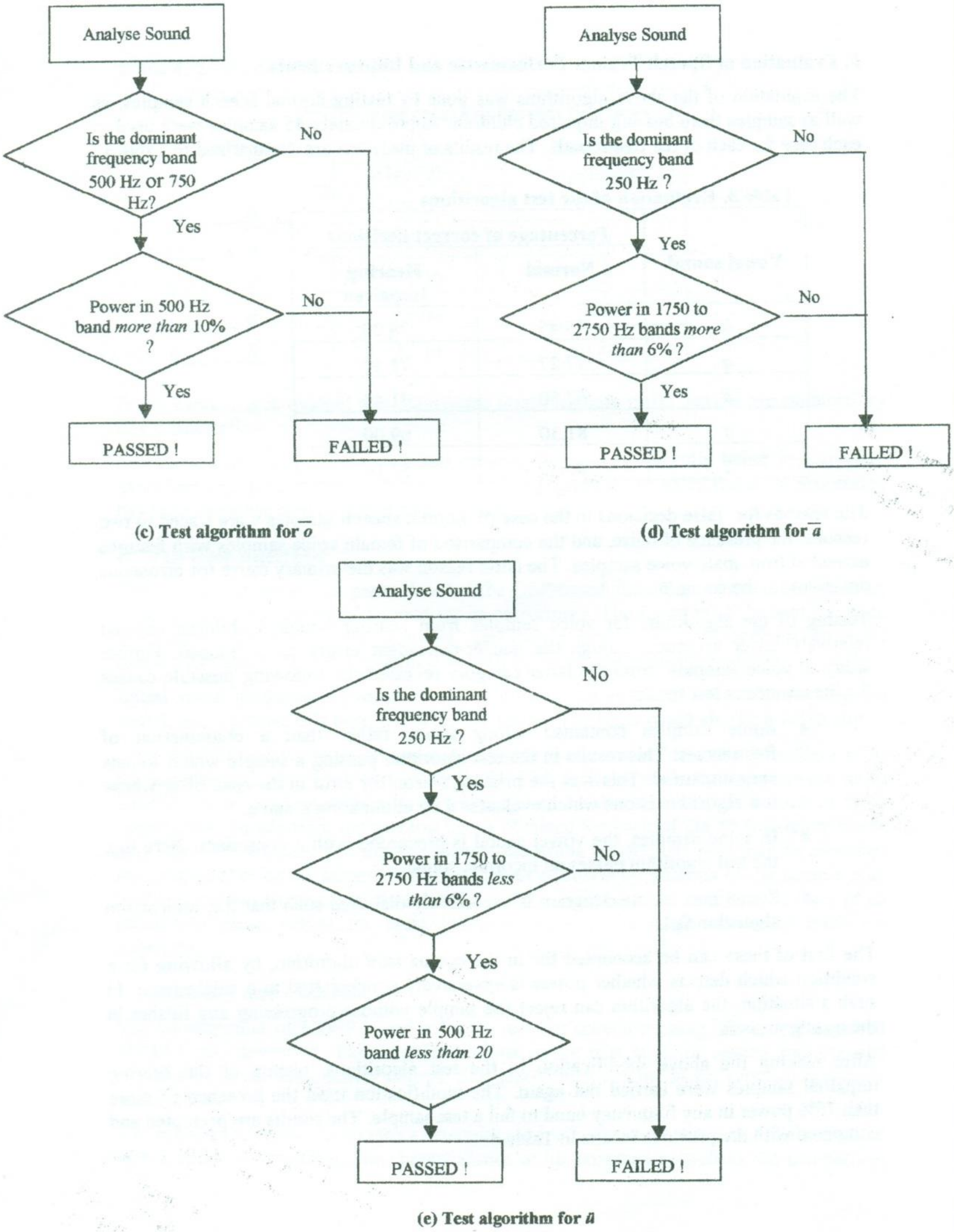
From the above features extracted through spectrographic anaylsis, the algorithms used to detect whether a sound has been articulated properly are developed. These algorithms are summarized in Figure 7 and are self-explanatory.


Analyse Sound
Is the dominant frequency band 500 Hz or 750 Hz?
No
Yes
Power in 1500 and 1750 Hz bands *less than* 10% ?
No
Yes
PASSED !
FAILED !
Analyse Sound
Is the dominant frequency band 250 Hz ?
No
Yes
Power in 1750 to 2500 Hz bands *greater than* 6% ?
No
Yes
PASSED !
FAILED !

**(a) Test algorithm for *ä***

**(b) Test algorithm for *a***

**Figure 7. Test algorithms for the vowel sounds**


Analyse Sound
Is the dominant frequency band 500 Hz or 750 Hz?
No
Yes
Power in 500 Hz band *more than* 10% ?
No
Yes
PASSED !
FAILED !
Analyse Sound
Is the dominant frequency band 250 Hz ?
No
Yes
Power in 1750 to 2750 Hz bands *more than* 6% ?
No
Yes
PASSED !
FAILED !
Analyse Sound
Is the dominant frequency band 250 Hz ?
No
Yes
Power in 1750 to 2750 Hz bands *less than* 6% ?
No
Yes
Power in 500 Hz band *less than 20* ?
PASSED !
FAILED !

**(c) Test algorithm for $\overline{o}$**

**(d) Test algorithm for $\overline{a}$**

**(e) Test algorithm for ă**

**Figure 7 contd. Test algorithms for the vowel sounds**

## 6. Evaluation of Speech Trainer Performance and Improvements

The evaluation of the above algorithms was done by testing normal speech samples, as well as samples from hearing impaired children. Approximately 45 samples were used in each case for each of the five sounds. The results of these test are summarized in Table 3.

**Table 3. Evaluation of the test algorithms**

| **Vowel sound** | **Percentage of correct decisions** | |
|---|---|---|
| | **Normal** | **Hearing Impaired** |
| *ä* | 95.45 | 78.00 |
| *a* | 77.27 | 71.40 |
| $\overline{o}$ | 67.50 | 71.40 |
| $\overline{a}$ | 81.30 | 69.00 |
| *ü* | 80.00 | 57.14 |

The reasons for false decisions in the case of normal speech samples were traced to two reasons: the presence of noise, and the comparison of female voice samples with features extracted from male voice samples. The latter reason was the primary cause for erroneous decisions in the sound $\overline{o}$.

Testing of the algorithms for voice samples from hearing impaired children showed relatively lower accuracy, though the gender-dependent errors were reduced. Further study of voice samples from the latter category revealed the following possible causes for the erroneous test results:

- Some samples contained strong tones, rather than a combination of frequencies. This results in the test algorithm passing a sample which sounds very unnatural. This was the primary reason for error in the case of *ü*, whose test algorithm is one which evaluates *ü* by eliminating $\overline{o}$ and $\overline{a}$.
- In some samples, the vowel sound is preceeded with a consonant. Here too, the test algorithm passes an incorrect sound.
- Sometimes the spectrogram is excessively distorted such that the comparison algorithm fails.

The first of these can be accounted for in the comparison algorithm, by allowing for a condition which detects whether power is excessively concentrated in a single band. In such a situation, the algorithm can reject the sample without progressing any further in the testing process.

After making the above modification to the test algorithms, testing of the hearing impaired samples were carried out again. The modification used the presence of more than 75% power in any frequency band to fail a test sample. The results are presented and compared with the previous results in Table 4.

**Table 4. Evaluation of modified test algorithms**

| Vowel sound | Percentage of correct decisions | |
|---|---|---|
| | **Original Algorithm** | **Modified Algorithm** |
| *ä* | 78.00 | 97.5 |
| *a* | 71.40 | 90.3 |
| *ō* | 71.40 | 89.5 |
| *ā* | 69.00 | 95.6 |
| *ü* | 57.14 | 90.24 |

From Table 4 it is evident that significant improvements in the results are obtained by this modification.

Elimination of errors due to an initial consonant sound is currently being investigated. Modifications to the autocorrelation of the spectrogram at the early stages of processing has been demonstrated to reduce these errors.

## 6. The User Interface

The above sections of the paper detailed the DSP-based functions of the speech trainer, which enables the testing of a sound for its correctness. The function of the user interface is to guide the user to utter a sound, and then provide him with feedback as to its correctness. Both the initial guidance, as well as the feedback must be provided visually.

Initial visual guidance is provided by a graphical display of the letter representing the sound, and a picture of a person making the sound. The latter could also be a video clip.

Once the user makes the sound and it is analysed by the trainer, the visual feedback as to its correctness is provided by two methods. The decision made by the software as to whether the sound was correct or not is conveyed by the traditional tick mark or cross respectively. In addition, the spectral levels obtained for each of the 16 frequency bands may also be displayed. The expected levels according to template data may be provided for reference, or as the target. The actual levels obtained by analysis of the sample may be displayed in a distinctly different manner. This could be integrated into the form of an interactive game, where the child learns to achieve the target levels by repetitive excercise.

## 7. Summary and Conclusions

The development of a DSP-based software tool for speech training of hearing impaired children was described. This is able to assist in the first stage of speech training: the articulation of the basic vowel sounds. The paper illustrates the signal processing aspects of the trainer.

The software analyses the trainee's utterance, and extracts characteristics to be compared with a template utterance. The characteristics to be extracted as well as the comparison algorithms were developed, and have been described in detail. The characteristics used for the detection of the correctness of the utterance are fairly simple, and are extracted using time-frequency analysis of speech samples. The comparison algorithms are based principally on the power distribution in the respective sounds.

The tool has been tested using samples of speech from hearing impaired children, and the results show success rates of approximately over 90% in the case of all vowels.

The user interface for the speech trainer, which is currently under development, uses visual feedback to indicate to the trainee, the degree of correctness of his utterance. It is intended to develop the user interface in the form of a game, where the child trains himself through repetitive attempts.

## 8. Acknowledgements

The authors whish extend their sincere appreciation to Sister Greta Nalawatte, of the Association for Hearing Impaired Children, Dalugama, Kelaniya, for introducing them to the need for software-based speech training. Her assistance regarding the basics of speech training proved to be invaluable.

The contributions of Mr. Janath Peiris to the work described in this paper are sincerely appreciated.

The authors also wish to thank students of the Electronic and Telecommunication Engineering Department, as well as those at the shool for hearing impaired, Dalugama, who assisted them in the collection of speech samples for testing.

## 9. References

[1] Matlab, *http://www.mathworks.com.*

[2] Matlab compiler. *http://www.mathtools.com*

[3] Mastering Visual C++, PBP Publications, India, 1997.

[4] J. W. Pitton, K. Wang and B-H Juang, "Time-Frequency Analysis and Auditory Modelling for Automatic Recognition of Speech", IEEE Communications Magazine, vol. 84, no. 9, pp. 1199 – 1215, September 1996.

[5] W. Koenig, H. Dunn, and L. Lacy, "The sound spectrograph", Journal of the Acoustical Society of America. Vol. 18, no, 1, pp. 19 - 49, 1946.

[6]R. Potter, G. Kopp, H. Green, "Visible Speech", New York: Van Nostrand, 1947.

[7] R. Rabiner, B. H. Huang, "Fundamentals of Speech Recognition", Englewood Cliffs, N.J.: Prentice Hall, 1993.

[8] H Dunn, "Methods of Measuring Vowel Formant Bandwidths", Journal of the Acoustical Society of America, vol. 33, no. 12, pp. 1737 – 1746, 1961.

[9] A. Holbrook and G. Fiarbanks, " Diphthong formants and their movements", Journal of Speech Hearing Research, vol. 5, no. 1. Pp. 38 – 58, 1962.

[10] Oppenheim, A. V., Shaefer, R.W.,"Discrete-Time Signal Processing", Prentice Hall Inc. Englewood Cliffs, New Jersey, 1989.