J.Natn.Sci.Foundation Sri Lanka 2024 52 (3): 283 - 298
DOI: http://dx.doi.org/10.4038/jnsfsr.v52i3.11717

RESEARCH ARTICLE

Electronics

Construction of an effective telephone bandwidth specifically for speaker recognition

T Thiruvaran
Department of Electrical and Electronic Engineering, Faculty of Engineering, University of Jaffna, Ariviyal Nagar, Killinochchi 44000, Sri Lanka.

Submitted: 22 May 2023; Revised: 20 March 2024; Accepted: 27 May 2024

Abstract: Although the telephone bandwidth is 0-4 kHz, the speaker specific information is not evenly distributed within this range but extends beyond 4 kHz. This is because the hypopharynx, especially the piriform fossa, affects the higher frequency region and contributes more to inter-speaker variation. By effectively shifting the higher bands to a lower frequency region, where speaker specific information is reduced, an effective 4 kHz bandwidth can be constructed to enhance speaker recognition performance. To achieve this a method was already proposed, which is extended in this paper to experimentally demonstrate and validate with more experiments. Furthermore, this paper defines the theoretically possible frequency space for which the frequency shifting method can be applied. To validate the method for different combinations of bands, possible bands were shifted in various directions in small steps. Speaker recognition experiments were conducted at each step to compare the performance against the baseband without any frequency shifting. Using the results of these extensive experiments, an approximate frequency space was defined where this frequency shifting performed better than the conventional baseband of 0-4 kHz signal. A simplified frequency shifting method was also investigated. Finally, the speech intelligibility of the frequency shifted narrow band speech signal was analyzed using objective speech quality measures. This showed that intelligibility was not significantly affected by the frequency shifting method.

Keywords: Frequency band shifting, speaker recognition, speaker specific information, speech intelligibility

INTRODUCTION

Automatic speaker recognition can be used as a remote biometric authentication when speech is transmitted over a telephone channel. The bandwidth of the speech signal that can be used for the authentication system is limited to 4 kHz so as to match the telephone bandwidth. This could be one of the reasons why the regular National Institute of Standards and Technology (NIST) speaker recognition evaluation is conducted using speech signals sampled at 8 kHz (NIST, 2008). On the other hand, it has been shown through several psycho-acoustic experiments that this narrow telephone bandwidth of 0-4 kHz is not the only frequency range that contains the most speaker specific information, but higher frequency regions have more speaker specific information (Dang & Honda, 1997; Kitamura *et al*., 2005; Lu & Dang, 2008; Hyon *et al.*, 2012).

Kitamura *et al.,* (2005) concluded that the hypopharynx behaves relatively stable during vowel sound production but varies more among different speakers. This conclusion was reached through a morphological analysis of inter-speaker and intra-speaker variation using MRI images. In terms of the frequency region, the spectra beyond 2.5 kHz are known to be

* Corresponding author (thiruvaran@eng.jfn.ac.lk; https://orcid.org/0000-0003-0569-9669)

affected by variations of the speaker's hypopharynx. The hypopharynx consists of the laryngeal tube and bilateral cavities of the piriform fossa, which are located at the bottom of the vocal tract. In particular, the piriform fossa showed inter-speaker variation in shape, area and depth of those cavities (Kitamura *et al.,* 2005).

Furthermore, Takemoto *et al.,* (2006) studied the role of the laryngeal cavity in relation to vocal tract resonance. The laryngeal cavity was found to contribute significantly to the formation of the fourth formant. It specifically suggests that the fourth formant is defined almost only by the geometry of the laryngeal cavity. In addition, the cluster of 3rd and 4th formants were also shown to be controlled by the geometry of the laryngeal cavity. This is a significant finding for speaker recognition since the geometry of the laryngeal cavity is a speaker dependent property. This again supports the fact that the higher formant region, compared to the lower formant region, is significant for speaker recognition. A recent analysis on identical twins, having implications on forensic speaker recognition, revealed that higher frequency formants have a greater speaker discriminative ability compared to lower frequency formants (Cavalcanti *et al.,*2021). In particular, the fourth formant was found to contain the highest speaker discriminative ability.

Further to the above findings of the lower vocal tract areas such as the piriform fossa that contribute to speaker specific information, Dang & Honda (1997) used MRI images of the piriform fossa to investigate their effect on the frequency spectra. The area function of the piriform fossa was shown to have more variation among different subjects but behaved relatively stable for vowel production. Further results revealed that the piriform fossa affected the frequency range from 4-5 kHz. This finding was again confirmed in an experiment using MRI based 3D printed vocal tract (Delvaux *et al.*, 2014). This further supports the findings that higher frequency regions are significant for speaker recognition.

In addition to the above morphological analysis, statistical based feature level analysis was done by Lu & Dang (2008) to quantify the distribution of speaker specific information in different frequency regions using the F-ratio, and mutual information on a 16 kHz database. It was concluded that three regions, specifically 100–300Hz, 4–5.5 kHz, and an area around 7.5 kHz contained more speaker specific information. Further, the study suggested that the area between 0.5 kHz and 3.5 kHz contained relatively less speaker relevant information. Their findings were validated by obtaining a better speaker recognition performance using a non-uniform filter bank that highlights the frequency regions with higher speaker specific information.

Another study analyzed the speaker specific information at the feature, model and classification level on NIST data (Thiruvaran *et al.*, 2009). At the feature level it used scatter matrix-based separation index measures. At the model level, it utilized a Kullback-Leibler distance-based measure. At the classification level, a series of speaker recognition experiments were conducted by leaving out one band to determine the contribution of that band on speaker recognition. The observations derived from these three levels of analysis for the telephone bandwidth data can be summarized as follows. The frequency range approximately between (1) 300-1000 Hz contributes to speaker recognition but not consistently in all three analyses; (2) 1000-2000 Hz contributes weakly to speaker recognition, consistent with Dang & Honda (1997); (3) 2000-3000 Hz contributes strongly to speaker recognition supporting the effect of the piriform fossa (Hyon *et al.*, 2012); and (4) 3000-3700 Hz is undetermined since it crosses the stop band region of the telephone bandwidth.

The study of a wide band signal with a bandwidth of 22.05 kHz reveals that the spectrum between 2000 - 5000 Hz contains significant speaker-specific information, as demonstrated in speaker recognition experiments (Thiruvaran *et al.*, 2009). Thus, all the above findings imply that the telephone bandwidth misses these speaker-specific information.

These observations provide motivation to develop a method that effectively utilizes the parts of speech bandwidth containing rich speaker-specific information. This method would enable the transmission over a telephone channel while retaining the advantage of remote authentication. The objective is to construct an effective 4 kHz band speech signal that includes speaker-specific information from both narrow and wide band signals. This signal will then be down sampled to an 8 kHz sampling frequency so that it can be transmitted over a telephone channel. To achieve this objective a frequency shifting method was proposed earlier (Thiruvaran *et al*., 2015). This paper seeks to extend this method by providing experimental validation for a range of frequencies. This shall be done with extensive experiments, reformulating the conditions, defining a theoretically possible frequency space for this method to be applicable, and finally studying the speech intelligibility of the frequency shifted signals. Based on the experiments an approximate frequency range will be determined where this method outperforms the baseband signal of 0-4 kHz. Additionally, a simple frequency shifting method without the lower band, would be shown to be ineffective as an alternative for the proposed frequency shifting method. Thereby this method will comprehensively demonstrate

the effectiveness of the frequency shifting method for speaker recognition.

This frequency-shifting method can be implemented through a mobile app that needs to be installed by clients or customers of a service institute, such as a bank or any institute, that requires secure access. The mobile app will convert the bandwidth of the speech signal using the proposed frequency-shifting method and then transmit it over the telephone channel.

It is worth noting that recently, methods for bandwidth extension for the recognition of narrow band speakers has been widely discussed (Abel & Fingscheidt, 2018; Nidadavolu *et al.* 2018; Kaminishi *et al.* 2019; Kamnishi *et al.* 2020). However, the method proposed in this paper is conceptually different from those bandwidth extension methods. Bandwidth extension methods involve adding extra bandwidth beyond the narrow band together with the full narrow band, resulting in an effective bandwidth of more than the narrow bandwidth of 4 kHz. On the contrary, the method discussed in this paper replaces some spectral content from the narrow band with some spectral content taken from beyond the narrow band while the total bandwidth remains at 4 kHz.

There is another context to the spectral shifting discussed in the literature that is related to cochlear implants (CI) where only a limited number of electrodes are used to process the speech. There are significant differences between the neural pattern generated by CI and the central speech pattern that post-lingually deafened patients previously developed in their brain (Fu *et al.*, 2005). One reason suggested is the spectral shift or distortion due to the CI electrodes compared to the natural cochlear. This spectral shifting is actually a distortion in the CI and the effect and implication of this spectral shifting is further discussed in literature (Li and Fu, 2010, Chennupati *et al.*, 2018). However, the use of the term 'spectral shifting' in the above CI related literature is entirely different from the one addressed in this paper. It is important to note that the spectral shifting discussed in this paper is not the same as in the aforementioned context.

MATERIALS AND METHODS

Frequency shifting process and conditions on band edge frequencies

This section briefly describes the process of shifting a high frequency band to replace a relatively less speaker-specific frequency band to form an effective telephone speech bandwidth, which was initially proposed in Thiruvaran *et al.*, (2015). Based on the above studies on the distribution of speaker-specific information, it can be broadly concluded that there are two distinct areas contributing significantly to speaker recognition compared to other areas of the spectrum. These areas are (i) in the very low frequency region, and (ii) extend beyond the narrow band telephone bandwidth region. These areas are symbolically represented in Figure 1 as band AB and band CD where the actual value of f_1, f_2 and f_3 would vary under different analysis. Hereafter, for simplicity of explanation, the first band and the second band are referred to as 'band AB' and 'band CD', respectively, as illustrated in Figure 1. The first band is loosely connected with the vocal fold and lower formant area, while the second band is related to the hypopharynx area in the human speech production system.

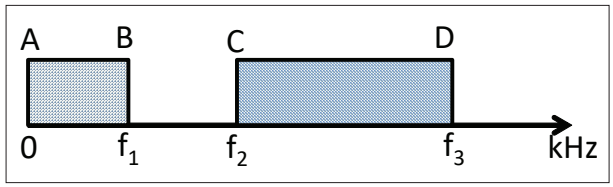

Figure 1: Frequency of bands contributing to speaker recognition (Thiruvaran *et al*., 2015)

If the biometric authentication is to support remote authentication, then the signal should be transmitted over the telephone channel with a bandwidth of 4 kHz. It is observed from the discussion in the introduction that the second band CD in Figure 1 spans outside the telephone bandwidth of 4 kHz. To effectively use the two bands, shown in Figure 1, a method is illustrated as a schematic diagram in Figure 2, which was first proposed in Thiruvaran *et al.*, (2009). First, the lower band AB is extracted using a low pass filter. Then the second band CD is extracted by a band pass filter. Subsequently, this band pass signal must be moved to fill the gap between f_1 and f_2 as shown in Figure 1. This moving or shifting is performed by modulating the band pass signal by a frequency of (f_2-f_1). Then this modulated signal is added to the lower band signal that is extracted by the first low pass filter. This is followed by applying a low pass filter to remove the content beyond the telephone bandwidth of 4 kHz. This filtering also serves as an anti-aliasing filter for the final down sampling step. After the down sampling, the shifted 4 kHz signal can be used for telephone transmission.

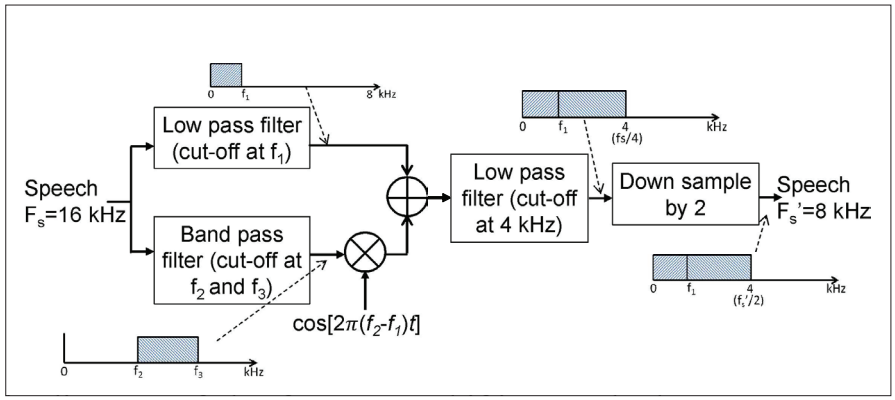

Figure 2: Schematic diagram of the frequency shifting method to construct an effective telephone bandwidth specific to speaker recognition. The frequency spectrums at different places in the schematic are also included (Thiruvaran *et al*., 2015).

Conditions on the band edge frequencies

The conditions to be satisfied by this frequency shifting method, which was previously derived (Thiruvaran *et al.*, 2015), are outlined here with modifications. The inequalities now involve only two parameters namely, f_1 and f_2. This adjustment aims to facilitate the subsequent experiments where these two parameters will be varied. In order to obtain a 4 kHz telephone bandwidth, the condition in (1) is introduced that ensures that band AB and band CD together form a 4 kHz bandwidth. The variables and values in all the equations and figures are in kHz.

$$f_3 - (f_2 - f_1) = 4 \quad \ldots(1)$$

Since this first condition (1) is an equation, the number of free parameters reduces from 3 to 2. Thus, the rest of the conditions are defined with two parameters, f_1 and f_2.

Further, it is necessary to avoid distortions due to modulation. There are two possibilities for distortions. The band after the bandpass filter before modulation is illustrated in Figure 3 (a). Let the magnitude spectrum after modulation be as illustrated in Figure 3 (b) where band C'D' and C''D'' are the modulated bands of the band pass signal of the band CD. The first possibility of distortion is when the frequency of point C'' becomes less than that of point D'; that is, in Figure 3 (b) if band C''D'' falls on to band C'D'. To avoid this distortion, equation (2) must be satisfied. When this condition is combined with equation (1), it reduces to (3), which is the second condition.

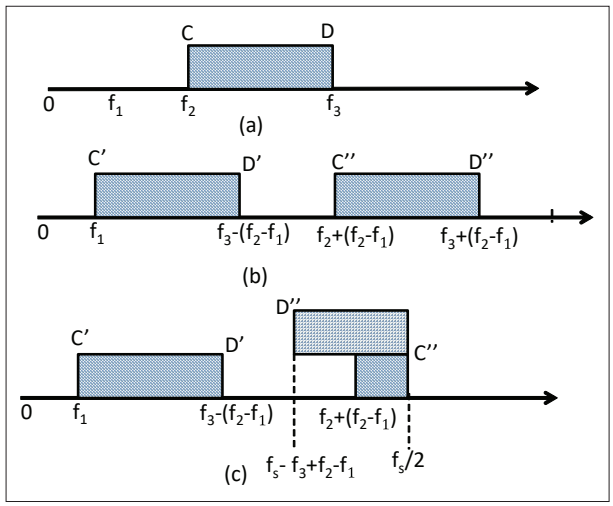

Figure 3: Positive side of the frequency spectrum after modulation. (a) spectrum of the band pass signal before modulation, (b) after modulation when there is no aliasing and (c) when there is aliasing; the aliased section is drawn with an opposite lines of hatching. (modified from Thiruvaran *et al*., 2015)

$$f_3 - (f_2 - f_1) \leq f_2 + (f_2 - f_1) \quad \ldots(2)$$

$$4 \leq 2f_2 - f_1 \quad \ldots(3)$$

The second possibility of distortion is when the band edge D'' is beyond $f_s/2$, where f_s is the sampling frequency of the original signal before down sampling. In this case the possibility of distortion arises when the band C''D'' is aliased onto the band C'D'. In Figure 3(c)

the point D'' has aliased since in this case $f_3 + f_2 - f_1$ is greater than $f_s/2$. The aliased point will fall back at $f_s - (f_2 + f_3 - f_1)$ as illustrated in Figure 3 (c). To avoid this distortion the point D'' should not fall onto band C'D'. For this, the condition in (4) should be satisfied. This condition reduces to (5) when f_3 is eliminated.

$$f_s - (f_2 + f_3 - f_1) \geq (f_3 - f_2 + f_1) \quad ...(4)$$

$$f_2 - f_1 \leq 4 \quad ...(5)$$

It should be noted that any aliasing of the band C''D'' on itself does not affect the final output since the subsequent low pass filter ultimately removes the band C''D''.

The motivation for this method is to include a higher band within the 4 kHz bandwidth. This means the first band AB should be less than 4 kHz, which is the fourth condition as in (6).

$$f_1 < 4 \quad ...(6)$$

Thus, there are four constraints on f_1, f_2 and f_3, namely equations (1), (3), (5) and (6) that must be satisfied for this frequency shifting method to be applicable.

This paper hypothesizes that by carefully selecting the frequencies, f_1, f_2 and f_3, the performance of the speaker recognition system can be improved compared to taking the band of 0-4 kHz when performing remote authentication through a telephone channel.

Theoretically possible frequency space for band edge frequencies

The theoretically possible frequency space suitable for the proposed frequency shifting is defined here using the conditions derived in the previous section when the sampling frequency was 16 kHz. Though the sampling frequency of the wide band signal considered is 16 kHz, this technique is applicable for any wide band signal for a sampling frequency higher than 16 kHz. However, the distribution of speaker specific information beyond 8 kHz is not that significant (Thiruvaran *et al.*, 2009).

The theoretically possible frequency range for bands AB and CD (Figure 1) is the range of possible band edge frequencies f_1, f_2, and f_3 that satisfy the four conditions defined in (1), (3), (5) and (6). Since equation (1) resulted in two free variables, f_1 and f_2 are taken as the free variables to define the theoretically possible frequency space for the band edge frequencies. This theoretically possible frequency space is shown in Figure 4 where the trapezoidal area defined by the lines QR, RS, ST, and TQ define the possible values of f_1 and f_2 that satisfy all the inequalities. Lines RS, QT, and ST represent the inequalities (3), (5) and (6), respectively.

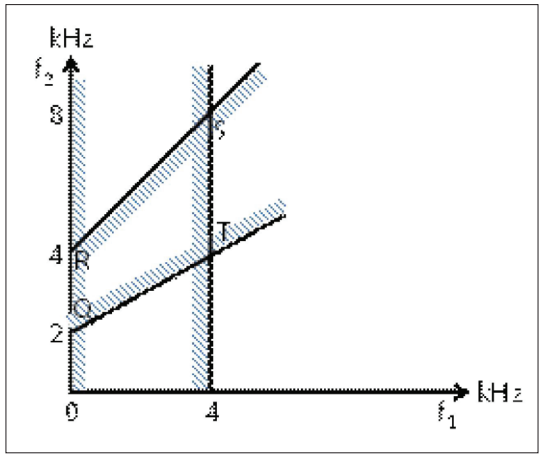

Figure 4: Theoretically possible frequency space for the frequency shifting method to be applicable in terms of the band edge frequencies of f_1 and f_2 for the bands AB and CD.

Selection of a suitable bandwidth for frequency band shifting

Experiment to determine the initial band AB

Initially, an empirical study was performed to check which frequency bands performed better after frequency shifting when compared to the conventional baseband signal of 0-4 kHz. That is, speaker recognition experiments were conducted with the signal obtained after performing this frequency shifting with different band edge frequencies. These results are compared with the conventional speaker recognition system that uses the 0-4 kHz bandwidth. Here the problem of selecting a suitable range of bandwidth for AB and CD (Figure 1) can be considered as selecting suitable values for the band edge frequencies f_1, f_2 and f_3 subject to the four conditions defined in (1), (3), (5) and (6).

Initially, speaker recognition experiments were conducted to determine the frequency area in the spectrum that contributes more to speaker recognition, i.e., to determine the spectral area that contains the most speaker discriminative information. This is studied by taking a single bandwidth of 1 kHz at a time and extracting the features for the speaker recognition experiment.

Experiments with varying bandwidth

The most suitable bands were found by gradually expanding or contracting bands AB and CD while ensuring that conditions defined by (1), (3), (5) and (6) were met. Since an empirical search spanning all possible scenarios would take a very long time, only a limited search was performed. For simplicity the starting point for the search for suitable bands was fixed by considering f_1 as 1 kHz, f_2 as 3 kHz and f_3 as 6 kHz.

Database and speaker recognition system

For this experiment a speech database of the IARPA Babel Tamil language collection sampled at 16 kHz was used (Chen *et al.*, 2015). Recordings of 247 male speakers in this database were used for this experiment with approximately 80 hours of recording. Each speaker had an average of about 300 recordings. Out of 247 speakers, 97 were used for the universal background model (UBM) and the rest 150 speakers used for evaluation. The evaluation list was prepared with 15,000 target and 89,400 non-target trials. The popular speaker recognition database of NIST, used for regular evaluations (NIST 2008), with a 8 kHz sampling frequency was not suitable for this experiment.

A Mel-frequency cepstral coefficient (MFCC)-based standard feature was used as the feature. An i-vector-probabilistic linear discriminant analysis (PLDA)-based backend was used (Ye *et al.*, 2016). The equal error rate (EER), which is the point where miss probability and false alarm probability are the same, was used as the performance metric.

Frequency shifting without the lower band – an alternative method

This section analyses a simpler modified method for frequency shifting by using a single band of 4 kHz bandwidth from a high frequency area beyond 2 kHz. The band EF in Figure 5 (a) is shifted according to the schematic diagram illustrated in Figure 5 (b). In this method, a single higher frequency band is obtained by a band-pass filter and then moved to the baseband by modulating it using an oscillator with a frequency of f_4. After modulation, the band that was shifted to the baseband is retained by the low-pass filter with a cut-off frequency of 4 kHz. This low pass filter will act as an anti-aliasing filter for the subsequent down sampling by a factor of 2, for this particular case, where the original sampling frequency was 16 kHz.

This system differs from that in Figure 2 by not having the low-pass filter parallel branch and the added module since there is no lower frequency band. This modification was motivated by the observation that the hypopharyngeal cavities are less affected by the phonetic content, which modify the spectrum beyond 2.5 kHz (Takemoto *et al.,* 2006).

Thereafter three-speaker recognition experiments were conducted by using a 4 kHz bandwidth of signal beginning after 2 kHz.

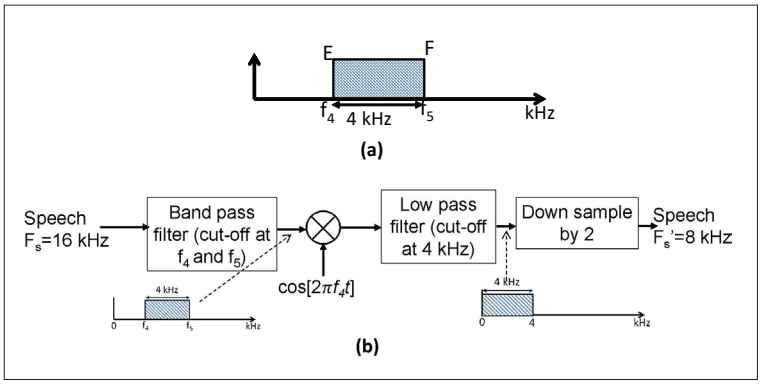

Figure 5: (a) The band EF of 4 kHz bandwidth in the higher speaker specific area as a single band, and (b) the schematic diagram of the frequency shifting to shift band EF to the baseband to form the narrow band 4 kHz signal.

Effect of frequency shifting on intelligibility

When the signal is distorted by the proposed method, the quality, specifically intelligibility, is an important factor to be analyzed. Though the proposed method was developed with the primary focus on speaker authentication, speech intelligibility cannot be compromised adversely. That is, the strength of the speaker specific information must be improved to increase the accuracy of the recognition of the speaker, as long as intelligibility is reasonably sufficient and not necessarily at its best. If the content of the speech is slightly unclear then it can be repeated to clarify the content.

In this section, the quality of the reconstructed speech signal, after the proposed frequency band shifting process, is objectively analyzed using the metrics perceptual evaluation of speech quality (PESQ) (International Telecommunication Union (ITU), 2001) and the virtual speech quality objective listener (ViSQOL) (Hines *et al*., 2015). ViSQOL gave comparable performance to perceptual objective listening quality analysis (POLQA), which is released by ITU as a licensed software (International Telecommunication Union, 2011). The purpose of this analysis is to determine if there is a significant drop in the intelligibility of the reconstructed signal.

One hundred recordings in the TIMIT database was used for this experiment. The original recordings of 16 kHz sampling frequency were converted to 8 kHz signals after performing the band shifting according to Figure 2. The recordings were then compared with the reference recordings obtained by directly down sampling (0-4 kHz baseband signal) without shifting the bands using PESQ and ViSQOL metrics. The reference recordings are the baseband signal used in speaker recognition experiments.

Filter design

Non-linear elliptic filters with 0.5 dB pass band ripple, 60 dB stop band attenuation, and a 500 Hz transition band were utilized in the filter design. Minor variations in the ripple and attenuation did not produce any significant change in the results. The elliptic filter was selected for its narrow transition band.

Phase information could not be preserved when a band outside the 0-4 kHz range is shifted within the band, regardless of whether a linear filter or non-linear filter was used. Non-linear filters are used to take advantage of lower computational complexity resulting from the use of lower order filters. Further, the phase distortion is tolerable in speaker recognition. The standard MFCC feature used in state-of-the-art speaker recognition systems does not preserve phase information but works well for that task. Though the phase is a concern in speech reconstruction, the purpose is not speech enhancement but to avoid complete loss of intelligibility. The quality of the reconstructed speech is always lower than the original since the process is optimized for speaker information and not for phonetic information. The band shifting is the primary reason for the drop in quality but the phase distortion due to non-linear filters may also cause a drop in the quality. The phase is distorted not only by non-linear filters but primarily by the band shifting process. Thus, even if a linear filter is used the phase distortion cannot be avoided. Further, as mentioned earlier, since the purpose is not speech enhancement the phase distortion is tolerable. Thus, non-linear filters were used.

RESULTS AND DISCUSSION

Results of speaker recognition experiments to find the initial band AB

For the experiments described to find initial band AB, the performance in terms of EER against the bandwidth is shown in Figure 6 where the x-axis shows the band used for each experiment. It can be assumed that a lower EER corresponds to a higher level of speaker specific information. The pattern of variation observed here is similar to the general pattern, discussed in the introduction, where the distribution of speaker specific information in the frequency space is discussed. It was observed that the 0-1 kHz band gave the lowest EER and thus can be considered as having the most speaker specific information. The 0-1 kHz band was chosen as the initial range for the band AB in the upcoming experiments in the next section. The objective was to test different band edge frequencies to determine the appropriate range of frequency for the above frequency shifting method. This observation partly justifies the selection of initial values of f_1 mentioned at the end of the experiments with varying bandwidths.

Further, it could be observed that the variation of the EER was not monotonic against the bandwidth. That is, the EER increases from 0-1 kHz to the 1-2 kHz band, and then decreases to the 3-4 kHz band, and then again increases thereafter. This shows that the speaker specific information in the bandwidth from 2-4 kHz is higher than that of the 1-2 kHz bandwidth.

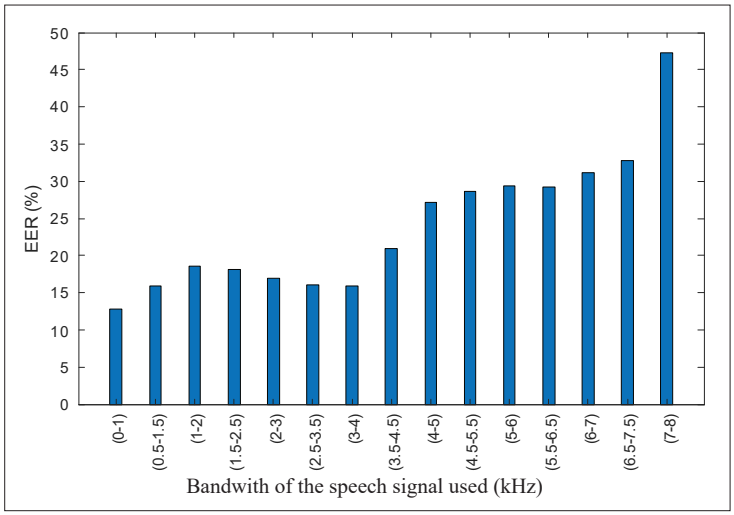

Figure 6: Speaker recognition performance of different areas of the frequency spectrum taking only 1 kHz of the speech spectrum in each experiment.

Results of experiments with varying bandwidths

The results of varying the frequencies of f_1, f_2 and f_3 as the bands are systematically changed, are shown and discussed for the following five cases.

Case I: Band AB is squeezed from point B and CD is expanded from point C

From the initial position, the band AB is reduced by changing f_1 from 1 to 0.2 kHz in steps of 0.1 kHz, while

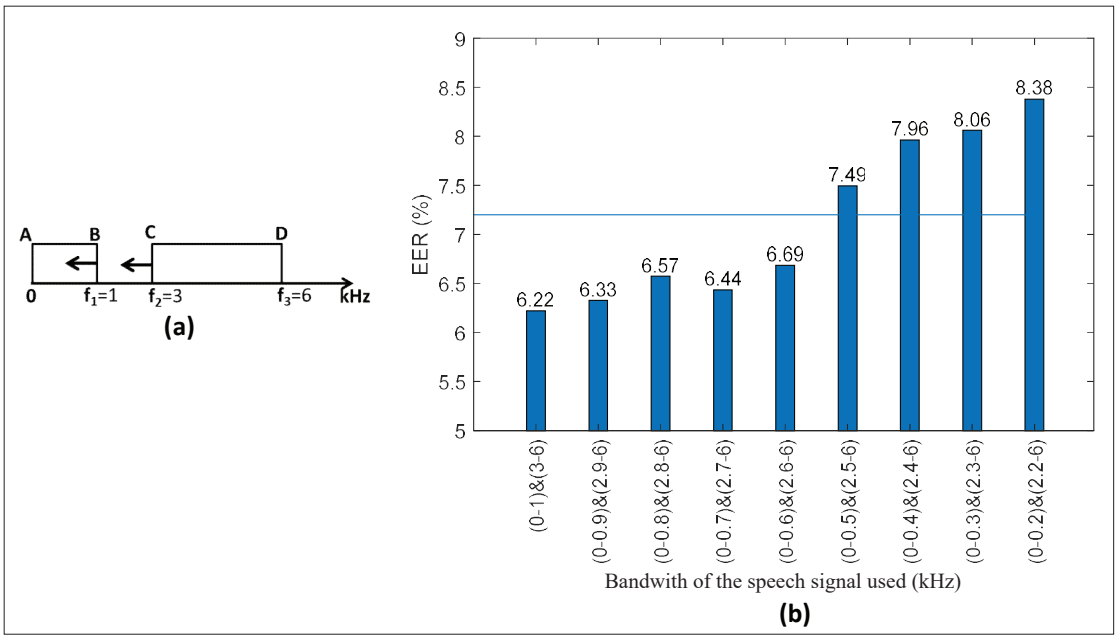

Figure 7: (a) The directions of bandwidth adjustments from the initial position in steps of 0.1 kHz, and (b) the speaker recognition performances. The performance when using the baseband of 0 to 4 kHz is superimposed as a line graph.

band CD is expanded by changing f_2 from 3 to 2.2 kHz in steps of 0.1 kHz (Figure 7a). This ensures that the total bandwidth of bands AB and band CD together will remain at 4 kHz, satisfying the requirements of the telephone bandwidth. In each case the bands were shifted and down sampled according to the schematic shown in Figure 2. The band shifted signal was used to extract features and speaker recognition experiments were conducted. The corresponding speaker recognition experiments with the step-by-step-variation for the 8 steps are shown in Figure 7 (b). Here the performance with the baseband (0-4 kHz band) is superimposed as a line. It can be observed that out of the nine cases experimented, five outperformed the results with the baseband of 1-4 kHz. When the band AB was reduced to 0.5 kHz or below, the performance dropped below the baseline system and thereby losing more speaker discriminative spectral contents. This supports the observation from Figure 6 where the 0-1 kHz band gave the lowest performance, which implied more useful information about the speaker was retained. The experiment with 0-1 kHz and 3-6 kHz will be replicated in all subsequent cases to ensure continuity.

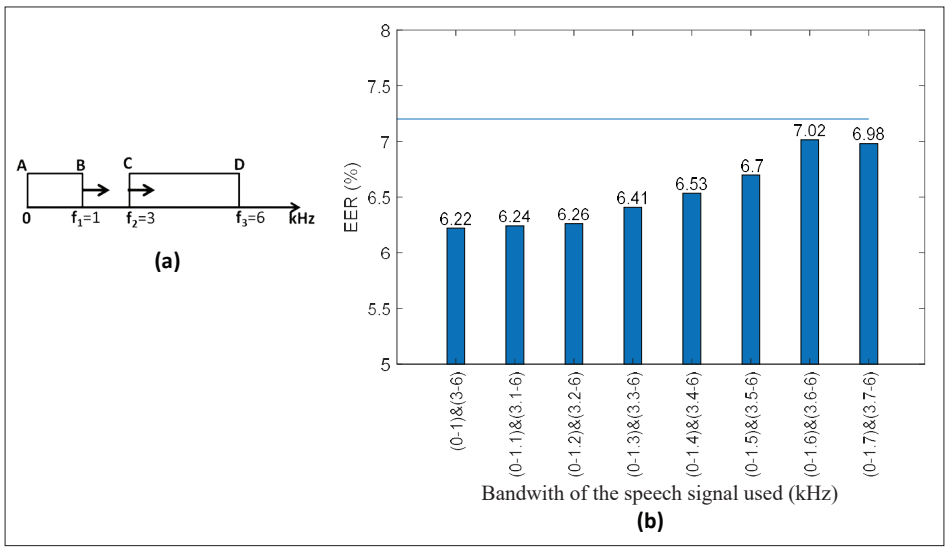

Figure 8: (a) The directions of bandwidth adjustments from the initial position in steps of 0.1 kHz at a time, and (b) the speaker recognition performances. The performance when using the baseband of 0 to 4 kHz is superimposed as a line graph.

Case II: B and AB is expanded from point B, and CD is squeezed from point C

Similarly, the band AB was expanded by changing f_1 from 1 kHz to 1.7 kHz in steps of 0.1 kHz while the band CD was reduced by changing f_2 from 3 kHz to 3.7 kHz in steps of 0.1 kHz (Figure 8a) and the corresponding speaker recognition results are shown in Figure 8b. The reason to stop at 1.7 kHz is to limit the number of experiments due to computational complexity. It can be observed that all the eight cases experimented have outperformed the results with a baseband of 1 to 4 kHz. The trend of increasing EER when the band AB is increased from 0-1 kHz to 0-1.7 kHz partly supports the observation in Figure 6 where the 1-2 kHz spectrum gave the lowest performance within the 4 kHz bandwidth, implying relatively lower speaker specific information.

Case III: Band AB is squeezed from point B and band CD is expanded from point D

In the same manner, the third possibility of varying the initial position of bands AB and CD is to change f_1 from 1 kHz to 0.2 kHz in steps of 0.1 kHz while band CD is expanded by changing f_3 from 6 kHz to 6.8 kHz in steps of 0.1 kHz (Figure 9a). The corresponding speaker recognition experiments are shown in Figure 9 (b) and the performance with the baseband (1 to 4 kHz) is superimposed as a line. Out of the nine experiments the results of five experiments outperformed the results with the baseband of 1-4 kHz. When the bandwidth of AB was reduced to 0.5 kHz or below, the performance did not improve more than the baseline system, similar to the observation in Figure 7.

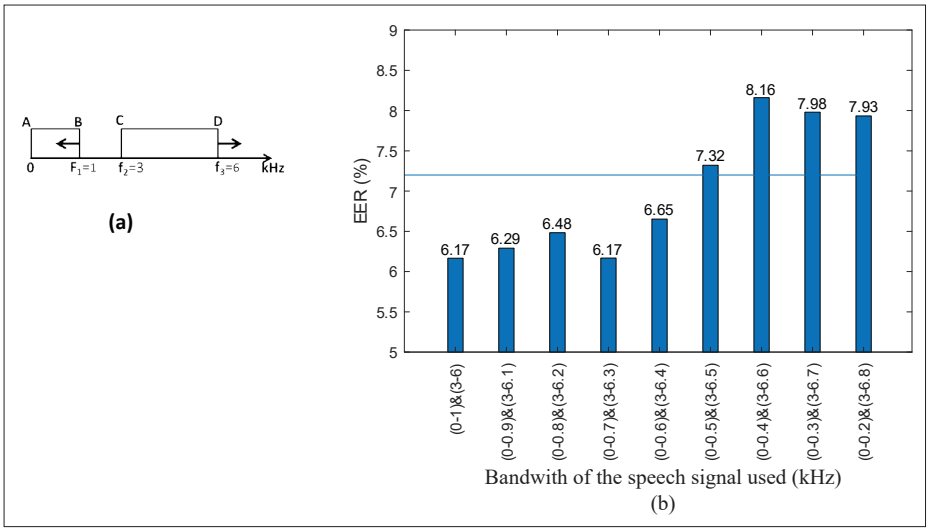

Figure 9: (a) The directions of bandwidth adjustments from the initial position in steps of 0.1 kHz, and (b) the speaker recognition performances. The performance when using the baseband of 0 to 4 kHz is superimposed as a line graph.

Case IV: Band AB is expanded from point B and band CD is squeezed from point D

As the fourth possibility for changing both bands, the band AB was expanded by changing f_1 from 1 kHz to 1.9 kHz in steps of 0.1 kHz while the band CD was reduced by changing f_3 from 6 kHz to 5.1 kHz in steps of 0.1 kHz. This is shown in Figure 10a and the corresponding speaker recognition results are shown in Figure 10b. It can be observed that all the experiments have outperformed the results with a baseband of 1 to 4 kHz.

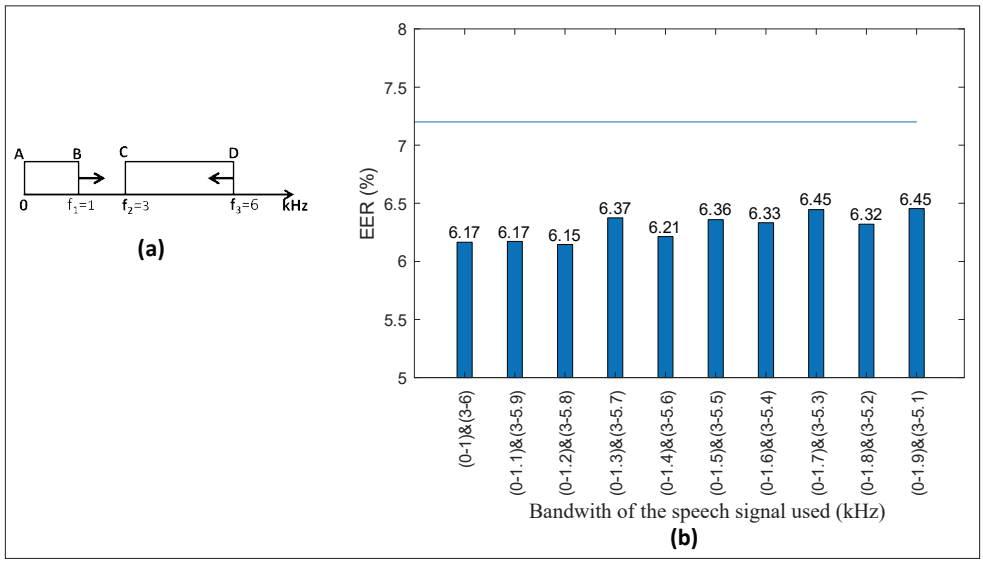

Figure 10: (a) The directions of bandwidth adjustments from the initial position in steps of 0.1 kHz, and (b) the speaker recognition performances. The performance when using the baseband of 0 to 4 kHz is superimposed as a line graph.

Case V: Band AB is fixed and band CD is shifted

As the final set of experiments, the band AB was fixed at 0 to 1 kHz but band CD was shifted. This is performed by changing f_2 from 2.5 kHz to 4 kHz while changing f_3 from 5.5 kHz to 7 kHz in steps of 0.5 kHz. Here the starting point of f_2 at 2.5 kHz (while f_1 still maintained at 1 kHz) touches one of the boundary lines QT as labelled in Figure 4; this implies f_2 cannot be reduced below 2.5 kHz while having f_1 at 1 kHz. Only four steps were tested including the previously tested band CD at 3 to 6 kHz; this is to streamline the number of experiments. This frequency change is illustrated in Figure 11a and the corresponding speaker recognition results are shown in Figure 11b. It can be observed that, out of the four experiments, three had outperformed the results with a baseband of 1 to 4 kHz.

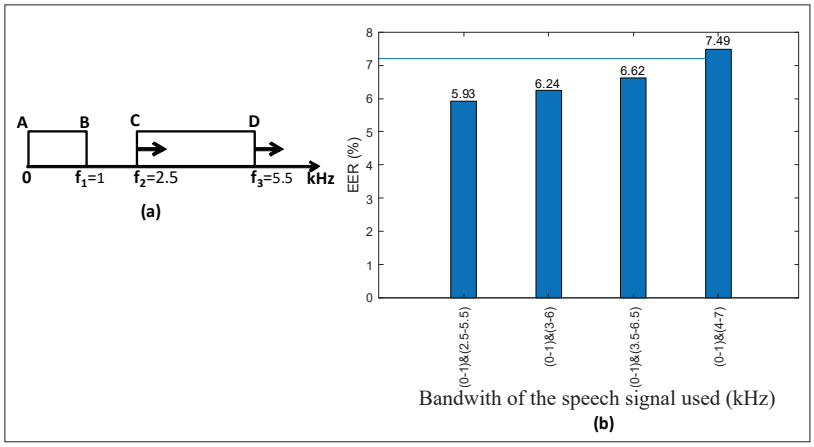

Figure 11: (a) The directions of bandwidth adjustments from the initial position in steps of 0.5 kHz, and (b) the speaker recognition performances. The performance when using the baseband of 0 to 4 kHz is superimposed as a line graph.

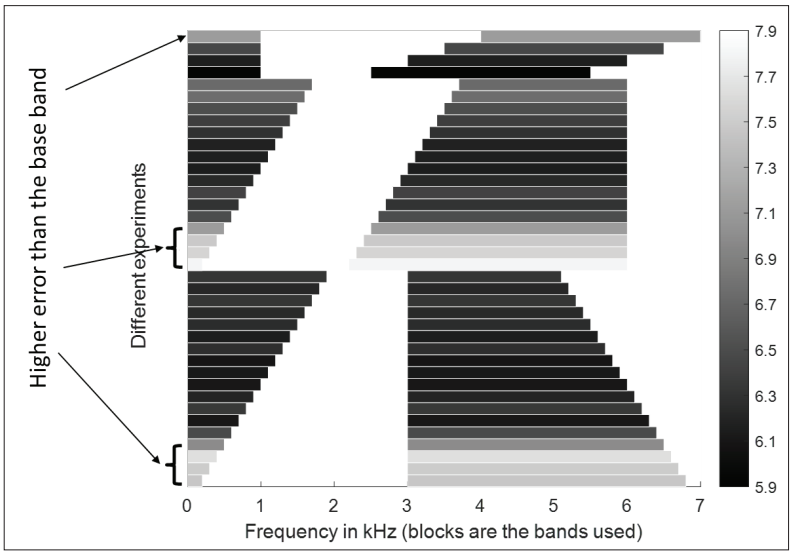

Figure 12: Summary of all the experiments performed under the 5 cases. The results of EER is coded gray and the corresponding bandwidths used are shown as blocks. The 9 experiments where the performance was not better than the baseband are also indicated.

The results of these five cases are summarised in Figure 12, with a gray colour code representing the corresponding EER, and the blocks represent the corresponding bandwidths. Considering all the five cases together it can be observed that out of 36 experiments with different bandwidths the proposed system performed better than the baseline system in 27 experiments. The nine cases where the performance was not better than the baseband are also indicated in Figure 12. Further a surface plot is shown in Figure 13 where the variation of EER is shown over the area spanned by f_1 and f_2. A plane is superimposed representing the EER when the baseband 0-4 kHz is used.

Considering all the above analysis the following conclusions can be made in a broad sense. When a part of the band (1 kHz) is taken from between 4 to 6.5 kHz (approximately), and replaced within the range 1.5 to 2.5 kHz region, the performance improved compared to using the baseband of 0 to 4 kHz alone. When the spectrum between approximately 0.5 to 1 kHz is not utilized, performance drops. Furthermore, the bandwidth of 0-1 kHz for band AB, and 2.5-3.5 kHz for band CD can be considered a local optimum. Overall, the results demonstrate the validity of the proposed method in improving the speaker recognition system.

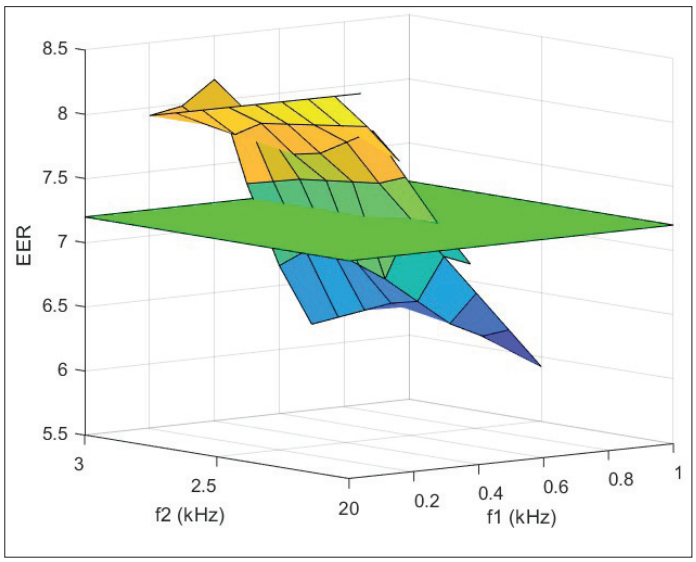

Figure 13: Surface plot showing the EER of all the experimented cases against f_1 and f_2. A plane is super imposed representing the EER when the baseband (0-4 kHz) is used.

Approximate suitable frequency space for band edge frequencies

This section superimposes the frequency points experimented in the previous sections under the five cases on the theoretically possible frequency space defined in Figure 4. The approximate suitable region is defined by connecting the end-points of the frequency points at which speaker recognition was better than that of the baseband signal. This approximate region is illustrated in Figure 14. The starting point ($f_1 = 1$, $f_2 = 3$ kHz) is named as point 'a', and the frequency points experimented in each case and the lines representing these experiments are as follows: case I – line ae, case II – line ab, case III – line af, case IV – line ac, case V – line ag and case VI – line ad. Since all the possible variations of f_1, f_2, and f_3 were not experimented, the suitable region found is an approximate area based on the limited experiments conducted.

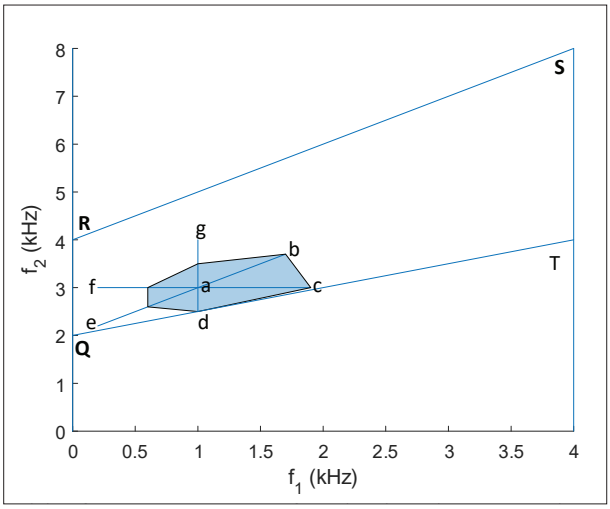

Figure 14: Approximate suitable frequency space for the bands AB and CD defined by the band edge frequencies of f1 and f2 superimposed on the possible frequency space.

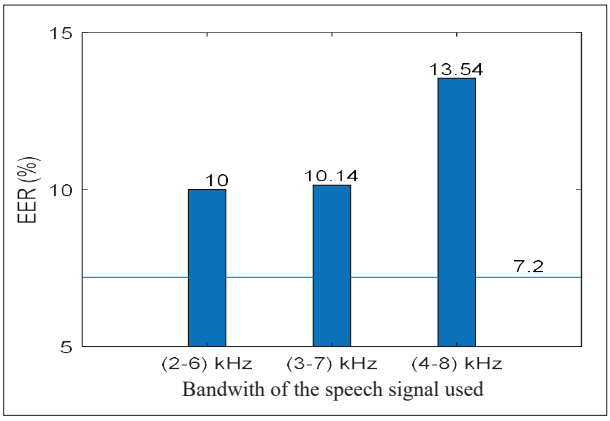

Figure 15: Speaker recognition results with shifted single band of 4 kHz bandwidth according to the schematic in Figure 15 (b), and the results of baseband (0 to 4 kHz) are superimposed as a line graph.

Results of the alternative method

The results for the three bandwidths chosen are shown in Figure 15 where the baseline results for the 0-4 kHz band is shown as a line graph. It can be observed that none of the experimented bands outperformed the baseline results i.e., shifting of a 4 kHz single band was not successful. Therefore, the band shifting method in Figure 2 was used. If this had performed better it would have facilitated a simpler system as illustrated in Figure 5 (b) than what is proposed in Figure 2. One possible reason why this did not perform better is that it does not use the low frequency region, possibly due to losing glottal related information. This again justifies the need to keep the lower frequency band and therefore the necessity for the added module in Figure 2.

Results of experiments on speech intelligibility

The box plot of PESQ scores and ViSQOL scores for the speech intelligibility experiments, for all the five cases discussed previously are shown in Figures 16 and 17, respectively, where each case starts with the bands of 0-1 kHz and 3-6 kHz, similar to the speaker recognition results given for different cases. The box plot shows the 2nd and 3rd quartiles as a box, mean as a triangle, median as a horizontal line, minimum and maximum as the ends of the lines and any outliers as circles.

When a threshold line is drawn for PESQ at 2 in Figure 16 the entire 2nd and 3rd quartiles are below the threshold line for the last four experiments in case I, and the last three experiments in case III. Interestingly, the last four experiments in case I and case III in the speaker recognition experiments also did not perform better than the baseband experiment as can be seen from Figures 7 and 9. That is, for the experiments where the speaker recognition performed better than that of the baseband, the PESQ score was also better. Similarly, for ViSQOL when a threshold line at 2 is drawn in Figure 17 the mean is below the threshold for the three experiments in case III and for the last experiment in case V. Here again the corresponding speaker recognition experiments did not perform better than the baseband signal as can be seen from Figures 7 and 11. Thus, for the cases where the band shifting performed better than the base band (the region shown in Figure 14), PESQ and ViSQOL gave a reasonable performance.

Further, the PESQ score is almost correlated with the bandwidth of band AB (Figure 17). For example, for case I, the bandwidth of band AB is decreased from 0-1 kHz to 0-0.2 kHz in steps of 0.1 kHz, and the PESQ score is monotonically reduced along that bandwidth change. A similar pattern can be observed for case III. For these two cases the speaker recognition performance is better than the baseband signal for the first 5 experiments (Figures 7 and 9) and the corresponding average PESQ for these 5 experiments were approximately higher than 2. That is, the bands that gave a lower PESQ also gave a higher EER in the speaker recognition experiments. ViSQOL almost follows the same trend for these two cases.

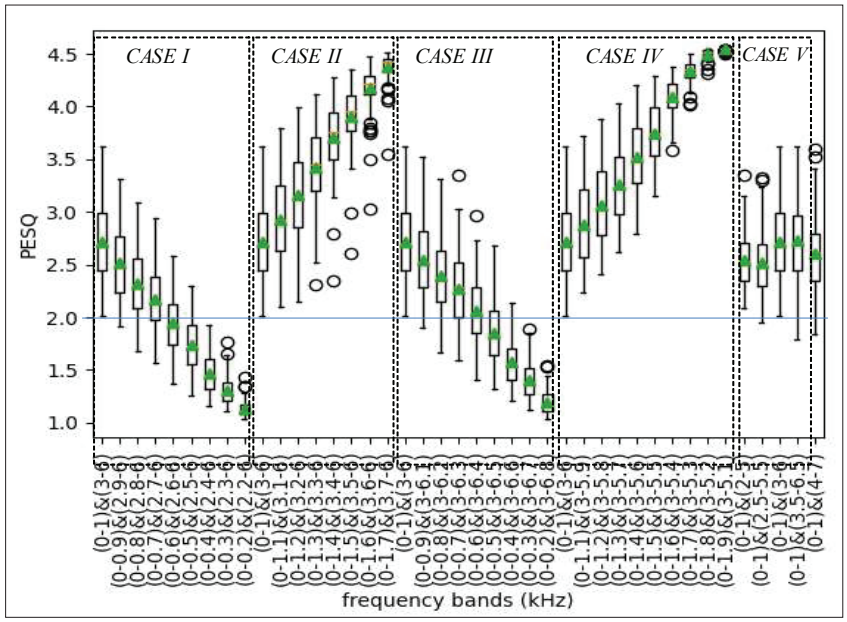

Figure 16: Box plots for PESQ score of the reconstructed signals using the proposed frequency band shifting method for 100 recordings from NTIMIT database against the frequency bands used. The cases are marked by a threshold line at drawn at PESQ 2.

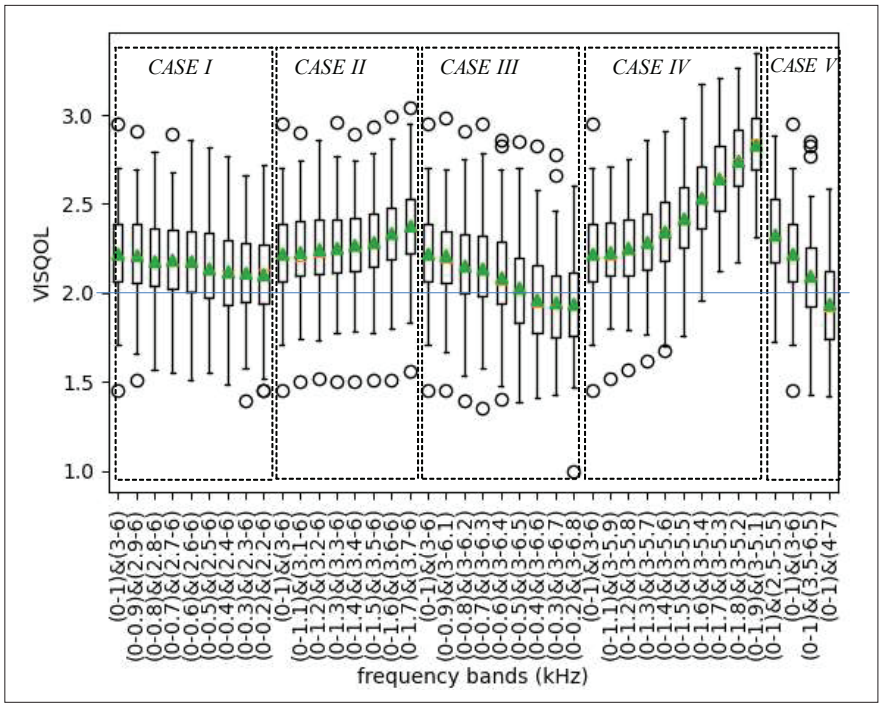

Figure 17: Box plot of ViSQOL score for the reconstructed signals using the frequency band shifting method for 100 recordings from the NTIMIT database against the frequency bands used. The experimental cases are marked by dash box and a super imposed threshold line at 2.

For case II and case IV, the bandwidth of band AB was gradually increased and the PESQ score also increased. A slight increasing trend was observed for the ViSQOL score as well. For these two cases the speaker recognition performance was better than the base band experiment for all the bandwidths experimented (Figures 8 and 10).

For case V where band AB was fixed at 0-1 kHz, the PESQ score shows a very small reduction when using the 2.5-5.5 kHz band compared to the 2-5 kHz band; it then slightly increases and then gradually declines, but overall remains unchanged. The corresponding speaker recognition performance showed a similar pattern but in the opposite direction, and the ViSQOL score declined gradually.

Overall, it can be concluded that the frequency shifting method when applied within the suitable region as depicted in Figure 14 does not significantly distort the signal. Additional informal listening tests have also shown that the quality of the reconstructed speech is sufficient to recognize the content of the speech.

Limited experiments on a dataset of a different language

A subset of the Voxceleb dataset was used to perform three experiments to determine whether the above findings are consistent when applied to a dataset of a different language. This subset had 200 speakers with 25,539 recordings in the English language. The selected three experiments were: (1) utilizing only the base band (0-4 kHz), (2) utilizing 0-1 kHz band and 2.5 – 5.5 kHz band, which yielded the optimal performance in the experiments outlined at the experiments with varying bandwidths, and (3) utilizing 0-1 kHz band and 3 – 6 kHz band. The EER obtained in these experiments were 8.2%, 7.05% and 7.5%, respectively. It is evident that both cases of band shifting (experiment 2 and 3) performed better than using only the baseband. Further, the band used in experiment 2 provided the best alignment with the findings from the experiments with varying bandwidths.

Limitations of the above experiments

The first limitation is the limited space beginning from point 'a' in Figure 14 for the experimental search due to the time-consuming nature of experiments. Even the variation from that point did not extend beyond a certain limit. This is because of the time-consuming nature of the speaker recognition experiments. The second limitation is that the experiments are not sufficient to find a global optimum. Since there are infinite different possible pairs of f_1 and f_2, it is difficult to experimentally find a global optimum. On the other hand, even if a global optimum is found this will have a bias based on the database used, back-end usage, and the parameters used. The third limitation is the use of an objective speech quality measure to study the intelligibility rather than a subjective speech quality measure since the purpose is only to verify if the intelligibility is retained and not to improve the listening quality. This is why an objective measure is preferred to the time-consuming subjective measure.

In addition to biometric authentication, speaker recognition is useful in areas such as in forensic applications and tracking communication by criminals for the law enforcing authority. For these and other applications it is necessary to use the 0-4 kHz bandwidth; the method proposed in this paper may not be applicable if the legal system challenges the use of a traditional bandwidth. However, for the application of remote biometric authentication, it could be possible to use an optimum bandwidth of 4 kHz that contains more speaker-specific information rather than only the 0-4 kHz band.

CONCLUSION

The frequency shifting method was empirically demonstrated for several different cases to perform better than only using the 0-4 kHz for narrow band speaker recognition. Further, for the four conditions required for the frequency shifting method, a theoretical possible frequency space is defined. Based on the experimental results, this space was further refined to find a suitable frequency space where the frequency shifting method performed better. A simple frequency shifting method without the lower band was demonstrated to be an inadequate method. Finally, the intelligibility of the frequency shifted signal was shown to be adequate to recognize the speech content. Based on the experiments the following conclusions can be made broadly. When a part of the band (1 kHz) is taken from approximately 4 to 6.5 kHz and replaced within the range of the 1.5 to 2.5 kHz region, the performance was improved compared to using only the baseband of 0 to 4 kHz band. When the spectrum approximately from 0.5 to 1 kHz is not used the performance declined. Furthermore, the bandwidth of 0-1 kHz for band AB, and 2.5-3.5 kHz for band CD can be considered as a local optimum. For future studies, the same frequency shifting method can be applied to other speech-related applications such as language recognition and emotion recognition by identifying the specific speech bandwidth containing the respective discriminative information.

Acknowledgements

The author would like to acknowledge the useful feedback from E. Ambikairajah and V. Sethu from the University of New South Wales, Australia

REFERENCES

Abel J. & Fingscheidt, T. (2018). Artificial speech bandwidth extension using deep neural networks for wideband spectral envelope estimation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing, 26*(1), 71-83. https://doi.org/10.1109/TASLP.2017.2761236

Cavalcanti, J. C., Eriksson, A. & Barbosa, P. A. (2021). Acoustic analysis of vowel formant frequencies in genetically-related and non-genetically related speakers with implications for forensic speaker comparison. *PLoS One, 16*(2), Article e0246645. https://doi.org/10.1371/journal.pone.0246645

Chen, N.F., Ni, C., & Chen, I. F. (2015, April 19-24). *Low-resource keyword search strategies for Tamil* (Conference abstract) IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), South Brisbane, Australia. https://doi.org/10.1109/ICASSP.2015.7178996

Chennupati, N., Kadiri, S. R., Yegnanarayana, B. (2018). Spectral and temporal manipulations of SFF envelopes for enhancement of speech intelligibility in noise. *Computer Speech and Language, 54,* 86-105. https://doi.org/10.1016/j.csl.2018.09.002

Dang, J. & Honda, K. (1997). Acoustic characteristics of the piriform fossa in models and humans. *The Journal of the Acoustical Society of America, 101*(1), 456-465. https://doi.org/10.1121/1.417990

Delvaux, B. & Howard, D. (2014). A new method to explore the spectral impact of the piriform fossae on the singing voice: Benchmarking Using MRI-Based 3D-Printed Vocal Tracts. *PLOS ONE, 9*(7), Article e102680. https://doi.org/10.1371/journal.pone.0102680

Fu, Q. J., Nogaki, G., Galvin III. J. J. (2005). Auditory training with spectrally shifted speech: Implications for cochlear Implant patient auditory rehabilitation. *Journal of the Association for Research in Otolaryngology, 6*(2), 180-189. https://doi.org/10.1007/s10162-005-5061-6

Hines, A., Skoglund, J., Kokaram, A. C. & Harte, N. (2015). ViSQOL: an objective speech quality model. *Journal on Audio, Speech, and Music Processing,* 2015, Article 13. https://doi.org/10.1186/s13636-015-0054-9

Hyon, S., Wang, H., Wei, J. & Dang, J. (2012). *A method of speaker identification based on phoneme mean F-ratio contribution* (Conference proceeding abstract). Annual conference of INTERSPEECH, Portland, USA. https://doi.org/10.21437/Interspeech.2012-349

International Telecommunication Union. (2001). *Perceptual evaluation of speech quality (PESQ), an objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs. ITU-T Recommendation*. https://www.itu.int/rec/T-REC-P.862

International Telecommunication Union. (2011). *Perceptual objective listening quality prediction*. *ITU-T Recommendation*. https://www.itu.int/rec/T-REC-P.863

Kaminishi, R., Miyamoto, H., Shiota, S. & Kiya, H. (2019). *Investigation on blind bandwidth extension with a non-linear function and its evaluation of x-vector-based speaker verification*. Twentieth annual conference of INTERSPEECH, Graz, Austria. https://doi.org/10.21437/Interspeech.2019-1510

Kamnishi, R., Miyamoto, H., Shiota, S. & Kiya, H. (2020). Blind bandwidth extension with a non-linear function and its evaluation on automatic speaker verification. *IEICE Transactions on Information and Systems*, 103(D), 42-49. https://doi.org/10.1587/transinf.2019MUP0008

Kitamura, T., Honda, K. & Takemoto, H. (2005). Individual variation of the hypopharyngeal cavities and its acoustic effects. *Acoustical Science and Technology, 26*(1), 16-26. https://doi.org/10.1250/ast.26.16

Li, T., & Fu, Q. J. (2010). Effects of spectral shifting on speech perception in noise. *Hearing Research 270*(1-2), 81-88. https://doi.org/10.1016/j.heares.2010.09.005

Lu, X., & Dang, J. (2008). An investigation of dependencies between frequency components and speaker characteristics for text-independent speaker identification. *Speech Communication, 50*(4), 312-322. https://doi.org/10.1016/j.specom.2007.10.005

Nidadavolu, P. S., Lai, C. I., Villalba, J. & Dehak, N. (2018). *Investigation on bandwidth extension for speaker recognition*. Nineteenth annual conference. INTERSPEECH, Hyderabad, India. https://doi.org/10.21437/Interspeech.2018-2394

NIST. 2008. *The NIST year 2008 speaker recognition evaluation plan*. http://www.nist.gov/speech/tests/sre/2008/index.html

Takemoto, H., Adachi, S., Kitamura, T., Mokhtari, P. & Honda, K. (2006). Acoustic roles of the laryngeal cavity in vocal tract resonance, *The Journal of the Acoustical Society of America, 120*(4), 2228-2238. https://doi.org/10.1121/1.2261270

Thiruvaran, T., Ambikairajah, E. & Epps, J. (2009). *Analysis of band structures for speaker-specific information in FM feature extraction*, Annual conference of INTERSPEECH, Brighton, UK. https://doi.org/10.21437/Interspeech.2009-39

Thiruvaran, T., Sethu, V., Ambikairajah, E. & Li, H. (2015). Spectral shifting of speaker-specific information for narrow band telephonic speaker recognition. *IET Electronics Letters, 51*(25), 2149-2151. https://doi.org/10.1049/el.2015.3117

Ye, J., Lee, K. A., Tang, Z. & *et al*. (2012). *PLDA modeling in I-vector and supervector space for speaker verification*. Thirteenth annual conference of INTERSPEECH, Portland, Oregon, USA. https://doi.org/10.21437/Interspeech.2012-460