Spectral shifting of speaker-specific information for narrow band telephonic speaker recognition

T. Thiruvaran✉, V. Sethu, E. Ambikairajah and H. Li

Speech-based authentication system can perform remote authentication over telephone channels. However, telephone channels are restricted to a bandwidth of ∼0–4 kHz while studies on the distribution of speaker-specific information in the speech spectrum strongly suggests that useful speaker-specific information is present above 4 kHz. A method to shift a part of this speaker-specific information above 4 kHz into the telephone bandwidth in place of less speaker-specific information originally present below 4 kHz is proposed. Speaker recognition experiments conducted using the proposed method leads to ∼18.5% relative improvement on equal error rate when compared to a system using the conventional telephone band speech, as evaluated on the Intelligence Advanced Research Projects Activity (IARPA) Babel Program Tamil language collection.

Introduction: One of the advantages of using speech as a biometric authentication is the ability to perform automatic remote authentication over a telephone line. This could be one of the reasons, presumably, why research on speaker recognition focuses more on 8 kHz sampled signals which is the sampling frequency used in telephone transmission. The National Institute of Standards and Technology (NIST) speaker recognition evaluation databases, which drive most research in speaker recognition, are primarily recorded over telephone channels sampled at 8 kHz. The telephone bandwidth of ∼0–4 kHz carries rich phonetic information and the primary purpose of telephone voice transmission is to convey the phonetic information. However, previous studies suggest that the 0–4 kHz band is not the optimum bandwidth for speaker recognition systems [1–3].

Statistical studies on the distribution of speaker-specific information across the frequency spectrum reveals that the higher frequency band beyond 4 kHz also contains rich speaker-specific information while a part of 0–4 kHz band carries relatively less speaker-specific information.

According to [1], the regions from 100 to 300 Hz, 4 to 5.5 kHz and 6.5 to 7.8 kHz contain significant speaker-specific information while the spectral region from 0.5 to 3.5 kHz was determined to contain limited speaker-specific information. An alternate study analysed speaker-specific information at feature, model and classification levels in speech sampled at 8 kHz and the results suggest that the spectral regions from 300 to 1000 Hz and 2.2 to 3.2 kHz contains significant speaker-specific information [2]. The same study also analysed speech sampled at 44 kHz and those results indicated that the spectral region from 2 to 5 kHz contained significant speaker-specific information.

A more recent study on speaker-specific information in different phonemes suggests that the distribution of speaker-specific information depends on the phonemes, but the trend of having a lower frequency region and a higher frequency region with significant speaker-specific information could still be observed [3]. On the basis these studies [1–3], it is somewhat safe to presume that significant speaker-specific information is present in the low frequency region below 1 kHz and in the higher frequency region above 2.5 kHz. Further, it can also be presumed that the region between 1 and 2.5 kHz contains limited speaker-specific information.

The above observation is further supported by some psychophysical research. Specifically, morphological analysis of MRI data on the influence of hypopharyngeal cavities revealed that the shape of the hypopharynx was relatively stable across different vowel production, but produced relatively large variation across different speakers [4]. Here hypopharynx includes cavities of laryngeal and piriform fossa. Thus, automatic speaker recognition system should focus on the bandwidth of speech spectrum where these hypopharynx has the influence. Further studies using the vocal tract transfer function of hypopharyngeal cavities revealed that the spectrum of speech beyond about 2.5 kHz is highly influenced by the cavities [4]. In addition, by studying the role of the laryngeal cavity in vocal tract resonance revealed that these cavities generate the fourth formant [5], partly supporting the previous observation. In [6], it is revealed that the effect of piriform fossa is in the region of 4–5 kHz in the speech spectrum. All these studies highly motivate to use the higher frequency spectrum for speaker recognition rather than just using 0–4 kHz bandwidth.

However, the use of spectral information above 4 kHz is not always possible in remote authentication systems given that telephone channels are bandlimited to ∼0–4 kHz. Consequently, it might be beneficial for speaker recognition systems if the spectrum of speech is modified so as to remove the spectral region that contains limited speaker-specific information and shift the higher frequency speaker-specific information into the telephone channel band.

This Letter proposes a specific technique to implement this spectral modification, designed to improve speaker recognition performance over telephone channels. In practice, this spectral shifting technique would probably be built into suitable telephones through which biometric authentication can take place.

Proposed spectral shifting technique: The proposed approach aims to preserve the speaker-specific information contained in the spectral regions highlighted in Fig. 1, i.e. 0 to f_1 and f_2 to f_3. This is accomplished by removing the content between f_1 and f_2 (does not contain significant speaker-specific information) and shifting the content between f_2 and f_3 such that all the speaker-specific information are contained within the 0–4 kHz band. It should be noted that f_1 and f_2 are ∼1 and 2.5 kHz, respectively, as previously mentioned, while f_3 is chosen such the following condition is satisfied in order to avoid any overlap during the proposed spectral shift:

$$f_3 - (f_2 - f_1) = 4\,\text{kHz} \tag{1}$$

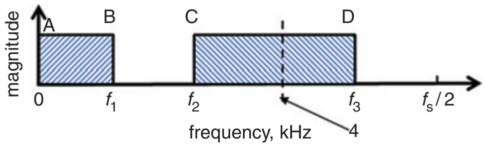

Fig. 1 *Spectral regions containing speaker-specific information that are to be preserved in transmission over telephone channels for speaker recognition*

The proposed technique for spectral re-construction is implemented in the time domain as outlined in Fig. 2. As indicated, the lower spectral band of interest (0 to f_1) is isolated using a lowpass filter and preserved. The higher spectral band (f_2 to f_3) is isolated using a bandpass filter and shifted to the region between f_1 and 4 kHz using an appropriate modulator.

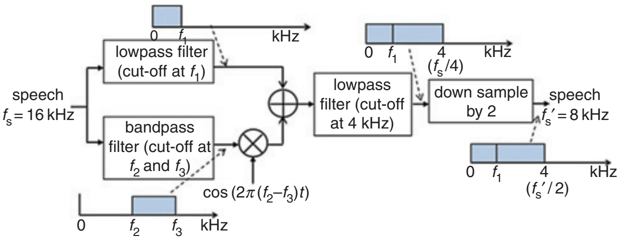

Fig. 2 *Schematic diagram of the proposed technique*

In the proposed scheme, there are two potential sources of spectral distortion that may arise due to the modulation-based spectral shift (see Fig. 3). Namely, if the spectral shift is too great the point denoted as C' (in Fig. 3) will fall below f_1 and the two spectral regions of interest will overlap. Alternatively, if the spectral shift is insufficient, the point indicated as D' will fall at a frequency greater than that at which the point C'' will fall and the two modulated copies of the shifted spectral region will overlap. The first potential source of overlap is avoided when the condition given by (1) is satisfied and the second potential source of overlap can be avoided if the condition given by (2) is satisfied

$$f_3 - (f_2 - f_1) \le f_2 + (f_2 - f_1) \tag{2}$$

Combining (1) and (2) leads to (3) below:

$$2f_2 - f_1 \ge 4\,\text{kHz} \tag{3}$$