Automated classification of vowel category and speaker type in the high-frequency spectrum

Jeremy J. Donai,[1] Saeid Motiian,[2] Gianfranco Doretto[2]

[1]Department of Communication Sciences and Disorders; [2]Lane Department of Computer Science and Electrical Engineering, West Virginia University, Morgantown, WV, USA

Abstract

The high-frequency region of vowel signals (above the third formant or F3) has received little research attention. Recent evidence, however, has documented the perceptual utility of high-frequency information in the speech signal above the traditional frequency bandwidth known to contain important cues for speech and speaker recognition. The purpose of this study was to determine if high-pass filtered vowels could be separated by vowel category and speaker type in a supervised learning framework. Mel frequency cepstral coefficients (MFCCs) were extracted from productions of six vowel categories produced by two male, two female, and two child speakers. Results revealed that the filtered vowels were well separated by vowel category and speaker type using MFCCs from the high-frequency spectrum. This demonstrates the presence of useful information for automated classification from the high-frequency region and is the first study to report findings of this nature in a supervised learning framework.

Introduction

Vowels are traditionally characterized by low-frequency (*i.e.,* below 3 kHz) spectral peaks, otherwise known as formants.[1] It is assumed that perceptual cues from the low-frequency region of the speech spectrum provide the requisite information for vowel identity, with less consideration given to the high-frequency region. Similarly, it is assumed that identification of speaker gender from an audio signal relies heavily on the low-frequency region of the speech spectrum.[2] Recent evidence reported by Donai and Paschall[3] found that normal-hearing listeners were capable of identifying vowel category from six vowel tokens produced by two male, two female, and two child speakers with substantial accuracy (*i.e.*, ranging from 40 to 70%) when low-frequency spectral peaks (at approximately 3.0-3.5 kHz) were removed via high-pass filtering. Classification results using linear discriminant function analyses from spectral peak information in the 3-12 kHz frequency region suggested that the signals were well-separated and inline with the perceptual data (in most cases within a few percentage points), thus supporting the perceptual findings. Additionally, Vitela *et al.*[4] examined vowel and consonant recognition from speech containing only high-frequency energy from 5.7 to 20 kHz produced by one male and one female speaker. Results revealed that normal-hearing listeners were able to extract information regarding vowel and consonant identity at performance significantly above chance, suggesting the presence of perceptual information in the high-frequency region of the speech signal.

Monson *et al.*[5] studied the ability of normal-hearing listeners to identify speaker gender using phrases from the Star Spangled Banner that were subjected to band-pass filtering at approximately 5.7 to 20 kHz. Results showed that listeners were able to use information from the high-frequency spectrum to identify speaker gender with accuracy significantly above chance (approximating 90% accuracy). In addition, Donai and Lass[6] recently reported gender recognition abilities among normal-hearing listeners using 250 ms vowel segments high-pass filtered at 3.5 kHz (above F3). Results indicated that normal-hearing listeners were able to identify speaker gender from signals comprised solely of high-frequency energy with greater than 80% accuracy. The authors suggested that the high-frequency region of the speech signal contained information that normal-hearing listeners could utilize to identify speaker gender with greater accuracy than would be predicted with low-frequency spectral information removed from the signal. Cumulatively, these studies provide documentation that human listeners are able to utilize information from the high-frequency region of the speech spectrum for perceptual tasks including vowel category and speaker gender judgments. These findings, however, have yet to be tested using an automated recognition framework.

The purpose of the following experiments is to determine the extent to which high-pass filtered vowel signals recorded in the Donai and Paschall[3] study could be separated using Mel frequency cepstral coefficients (MFCCs) in a supervised learning approach. MFCCs incorporate aspects of audition and are a popular method utilized in automatic speech recognition systems.[7] Given the previously reported behavioral data, it is hypothesized that vowel signals comprised solely of high-fre-

Correspondence: Jeremy J. Donai, West Virginia University, 355 Oakland St., Morgantown WV 26506, USA.
Tel.: +1.304.293.2662 - Fax: +1.304.293.2905.
E-mail: jeremy.donai@mail.wvu.edu

Key words: Classification; formants; high-frequency; mel frequency cepstral coefficients; vowels.

Acknowledgements: thank you to Rachel Halbritter for her assistance with manuscript and stimuli preparation.

Contributions: the authors contributed equally.

Conflict of interest: the authors declare no potential conflict of interest.

Received for publication: 1 June 2015.
Revision received: 30 December 2015.
Accepted for publication: 7 January 2016.

Audiology Research 2016;6:137
doi:10.4081/audiores.2016.137

quency energy will provide useful information for determining vowel category and speaker type in an automated recognition task.

Materials and Methods

Dataset

Vowels were recorded from two males, two females, and two children (one male and one female, both age 10) at 96 kHz and 24-bit resolution in an h/Vowel/d (hVd) context using a high fidelity microphone (Lawson 251) as described in Donai and Paschall.[3] Stimuli for the experiment consisted of five productions of six naturally produced, high-pass filtered hVd signals; /æ/ as in *had*, /i/ as in *heed*, /u/ as in *who'd*, /ɔ/ as in *hawd*, /ɜ/ as in *herd*, and /e/ as in *hayed*. The hVd signals were recorded in connected speech using the carrier phrase, *I say (hVd) again*. The individual hVd signals were extracted from the audio file for processing. In addition, a 100 ms segment from the central portion of the vowel was visually calculated, extracted, and windowed using a 10 ms hanning window with 50% overlap. This was done to evaluate classification from a more steady-state portion of the vowel, thereby reducing the potential effects of co-articulation in the hVd productions. The full and segmented signals were then down sampled to 48 kHz and saved for analysis. The signals were high-pass filtered to remove spectral peak information at and below the third formant (F3), which varied the filter cutoff used for the male, female, and child speakers. The vowels were filtered using an equiripple finite impulse response filter. The male signals were filtered using a 3.0 kHz stopband and 3.1 kHz passband; the females were filtered using a 3.1 kHz stopband and 3.2 kHz passband; the child signals were filtered using a 3.4 kHz stopband and 3.5 kHz passband. All filters were designed to provide 120 dB of attenuation, with 1 dB passband ripple.

Classification approach

The classification is based on a number of steps. First, MFCCs are extracted from each vowel signal, thus forming a temporal sequence. This is modeled as the output of a dynamical system. Therefore, classification requires the ability to classify dynamical systems, and this in turn entails having a way for comparing them. Among various metrics for computing the similarity between dynamical systems, the family of Binet-Cauchy kernels has been shown to have good performance in different applications.[8,9] Those kernels are then used for training a support vector machine (SVM) classifier in a supervised learning framework.[10]

MFCCs represent audio features computed following a signal-processing pipeline motivated by perceptual and computational considerations.[11] They are extracted from the Mel-frequency cepstrum, which is a representation of the short-term power spectrum of sound. More pre-

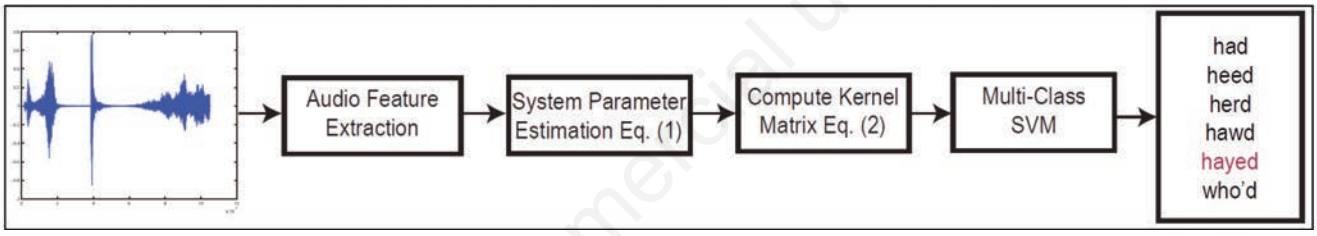

Figure 1. Visual illustration of the processing steps used in the classification experiments.

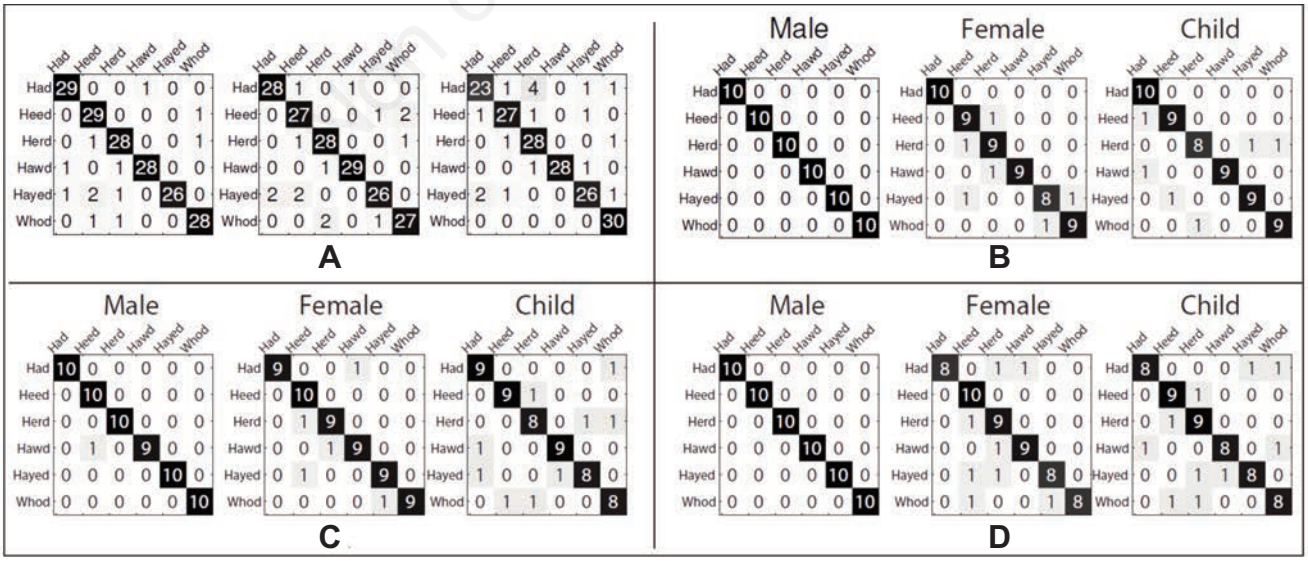

Figure 2. A) Confusion matrices for the first experiment. Full set with 96 kHz rate (left). Full set with 48 kHz rate (middle). Segmented set with 48 kHz rate (right). B-D) Confusion matrices for the second experiment: B) Full set with 96 kHz rate; C) Full set with 48 kHz rate; D) Segmented set with 48 kHz rate. Note: Vowel category actual on the vertical axis and vowel classified on the horizontal axis.

cisely, the audio signal is first divided into frames using a windowing function, and the power spectrum of each frame is computed. Then, a Mel filterbank is used to compute the signal energy in various frequency regions. Since human auditory sensitivity is non-linear, the logarithm of the energies is taken. Finally, the discrete cosine transform of the log-filtered energies is used to compute uncorrelated energy coefficients. Figure 1 provides a visual of the processing steps used in the current experiments.

Since MFCCs lie on a non-Euclidean space,[12] indicated here with $\mathcal{S}$, the model of choice for this work is a non-linear dynamical system (NLDS). More precisely, given a temporal sequence $\{y_t\}_{t=1}^{T}$, representing a vowel signal, y_t is an MFCCs feature vector that lies on $\mathcal{S}$. Each $y_t \in \mathcal{S}$ is then mapped to $\mathcal{H}$, a Reproducing Kernel Hilbert Space, using a mapping function $\Phi : \mathcal{S} \rightarrow \mathcal{H}$, and its temporal evolution is modeled as the output of a linear dynamical system (LDS), given by Eq. (1).

$$\begin{cases} x_{t+1} = Ax_t + v_t, \\ \Phi(y_t) = Cx_t + w_t. \end{cases} \quad (1)$$

Here $x_t \in \mathbb{R}^n$ is the state at time t, $A \in \mathbb{R}^{n \times n}$ describes the dynamics of the state evolution, the system noise v_t and the observation noise w_t are assumed to be independent zero-mean i.i.d. Gaussian distributed, and C is a linear operator $C : \mathbb{R}^n \rightarrow \mathcal{H}$.

A family of Binet-Cauchy kernels for LDSs has been extended for comparing NLDSs of the type in Eq. (1).[9] In particular, the Binet-Cauchy trace kernel for NLDS is the expected value of an infinite series of weighted inner products between the outputs after embedding them into the high-dimensional (possibly infinite) space $\mathcal{H}$, using the map $\Phi(\cdot)$. More precisely described in Eq. (2).

$$K(\{y_t\}_{t=1}^{\infty}, \{y'_t\}_{t=1}^{\infty}) \doteq E\left[\sum_{t=1}^{\infty} \lambda^t \Phi(y_t)^{\mathrm{T}} \Phi(y'_t)\right] = E\left[\sum_{t=1}^{\infty} \lambda^t \kappa(y_t, y'_t)\right] \quad (2)$$

where $0 < \lambda < 1$ and $\kappa(y_t, y'_t) = \Phi(y_t)^{\mathrm{T}} \Phi(y'_t)$ is a Mercer kernel.

Given the vowel signal $\{y_t\}_{t=1}^{T}$, the parameters of model (1) can be estimated as explained by Chaudhry *et al.*[9] Provided the covariances of the system noise, the observation noise, the dynamics of the state evolution, and the initial state, kernel (2) can be computed in closed form.[9] Considering the non-linear nature of $\mathcal{S}$, the kernel $\kappa(\cdot,\cdot)$ is chosen to be a Gaussian RBF kernel with Euclidean distance. Finally, classification is performed by training a one-*versus*-all SVM multiclass classifier, using the LIBSVM package.[10]

Leave-one-out cross-validation was used to compute the classification accuracy and the parameter of the RBF kernel κ while λ was set to 0.9. For the first set of the dataset, we used 96 kHz and 48 kHz sampling rates to compute MFCCs. It was of interest to determine the effect, if any, sampling frequency had on classification accuracy, using a more traditional sampling frequency (48 kHz) and a less-commonly used, and higher, sampling frequency (96 kHz) used in Donai and Paschall.[3] For the second set, a 100 ms central portion of the vowel at a 48 kHz sample rate was used. To compute 13 dimensional MFCCs, window length and step time were set to 20 ms and 3 ms, respectively.

Results

Four experiments were designed to determine the extent to which the six high-pass filtered vowel signals could be separated despite their within-class variability. The first experiment combined all vowel samples (male, female, and child), producing 6 vowel classes, with 30 samples per class (vowel). In the second experiment, the same dataset was used, but in this case the classification accuracy for the male, female, and child vowels were computed separately. The third experiment was designed to classify the filtered vowels by speaker types. All vowel samples are combined to create 3 classes (male, female, and child), with 60 samples per class. Lastly, in the fourth experiment, classification of speaker type was performed by individual vowel category. For the purpose of this study, the signals produced by the two children of different genders (both age 10) were combined because they were both reportedly pre-pubescent and observed to be of similar stature on the day of the recording.

Table 1 shows the classification accuracies for the first and second experiment, with Figure 2A and Figure 2B-D providing confusion matrices, respectively. As shown in Table 1, cumulative classification accuracy of vowel category ranged from 83 to 100% for both sample rates and for the full hVd and segmented signals (for specific error patterns refer to Figure 2A-D. Additionally, as shown in the first row of Table 2, cumulative classification of speaker type (male, female, or child) ranged from 85 to 91% (for specific error patterns refer to Figure 3), with classification accuracy rates above 83% for each speaker type in each condition (full *vs* segmented and for both sampling frequencies).

Results of the fourth experiment are displayed at the bottom of Table 2. When classifying speaker type by individual vowel category, performance was perfect for all vowels with the exception of /u/ (who'd) produced by child speakers. In each of the few misclassifications, the child vowel

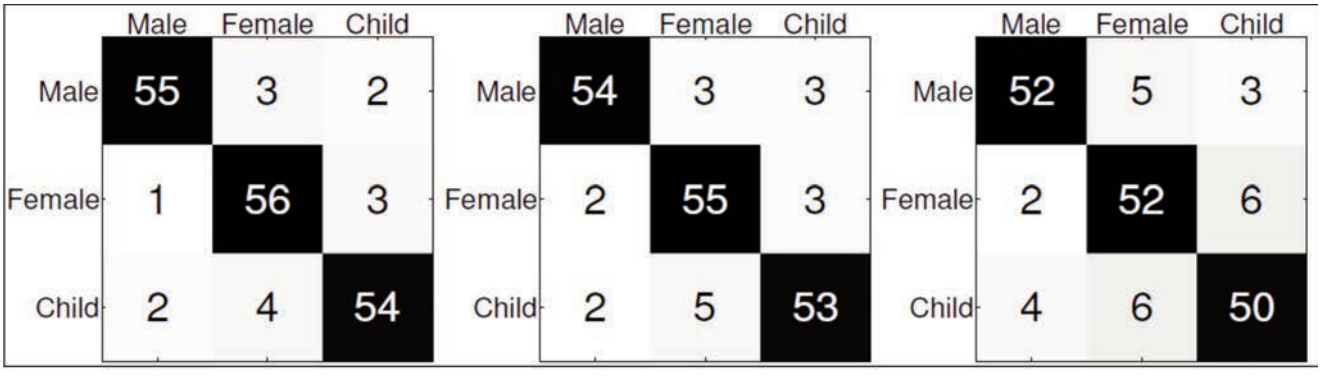

Figure 3. Confusion matrices for the third experiment. Full set with 96 kHz rate (left). Full set with 48 kHz rate (middle). Segmented with 48 kHz rate (right). Note: Speaker type actual on the vertical axis and speaker type classified on the horizontal axis.

was classified as being produced by a female speaker (confusion matrices not shown because of very few observed misclassifications). As such, specifying vowel category in the fourth experiment improved cumulative classification of speaker type by approximately 8 to 13% compared to experiment three where this signal attribute was not specified during classification. Cumulatively, results of the four experiments described here represent an initial, and successful, attempt to classify various aspects of the speech signal when relying solely on high-frequency features.

Discussion

With a dataset containing multiple productions of vowels from two male, two female, and two children, classification results suggest the high-frequency region to contain useable information for classification of vowel category and speaker type in a supervised learning framework. Classification accuracy rates for individual vowel category for the full set of hVd signals (combined male, female, and child signals used in the first experiment) ranged from 77 to 100% (with cumulative performance ranging from 90 to 93%), suggesting good classification performance when using a combined set of high-pass filtered vowel productions from three different speaker types. Identification accuracy of this order was found for the full hVd signals at 48 and 96 kHz sample rate as well as the 100 ms segments at a 48 kHz sample rate (see results of the first experiment in Table 1). Additionally, results showed good performance (*i.e.*, cumulative accuracy rates from 85 to 91%) for classifying speaker type (male, female, or child) from the high-pass filtered vowels (Table 2). Taken together, these findings suggest the existence of usable features from the high-frequency portion of the speech spectrum (above F3). The current results are novel in that classification of individual sound and speaker type solely from high-frequency information, until now, was yet to be described in the literature.

The cumulative outcomes from these experiments have potential implications for automated speech/speaker recognition performance, particularly in degraded acoustic environments containing substantial low-frequency spectral energy. In these conditions, the use of high-frequency information from the vowel signal has the potential to provide additional information to improve classification accuracy when the low-frequency spectrum is obscured by noise (*i.e.*, conditions with missing or unreliable data).

Earlier research by Hayakawa and Itakura[13] investigated the effects of extending the signal bandwidth for speaker recognition systems in quiet and in noisy conditions and found benefit (*i.e.*, reduced error rates in noise) when including additional high-frequency information up to 10 kHz from the speech signal. More recently, Hu and Wang[14] investigated the benefits of including information from the high-frequency region for segregating speech from a speech-in-noise mixture and found improved performance (average increase of 5.2 dB signal-to-noise ratio compared to previous versions) when including features from the high-frequency portion of the signal. Additionally, Deshpande and Holambe[15] investigated individual speaker recognition from sentences in the presence of competing vehicle noise and reported substantial information regarding individual speaker identity in the high-frequency band above 4 kHz, with speaker identification accuracy over 90% using features in the 4-8 kHz region. It should be noted that differences in the types of speech signals used for classification in the aforementioned studies and the current study exist; however, what is common amongst them is that in each case the use of information from the high-frequency region of the speech signal was useful for classification purposes, which makes them relevant to the current experiments.

Table 1. Vowel category classification accuracy for the first and second experiment. Full set refers to the full hVd signals while segmented refers to the steady state portion of the vowel extracted from the full hVd signal.

	First experiment			Second experiment								
	Full set 96 kHz	Full set 48 kHz	Segmented set 48 kHz	Full set 96 kHz			Full set 48 kHz			Segmented set 48 kHz		
Vowels				Child	Female	Male	Child	Female	Male	Child	Female	Male
had	96.67	93.33	76.67	100	100	100	90	90	100	80	80	100
heed	96.67	90	90	90	90	100	90	100	100	90	100	100
herd	93.33	93.33	93.33	80	90	100	80	90	100	90	90	100
hawd	93.33	96.66	93.33	90	90	100	90	90	100	80	90	100
hayed	86.67	86.67	86.67	90	80	100	80	90	90	80	80	100
who'd	93.33	90	100	90	90	100	80	90	100	80	80	100
Cumulative	93.33	91.67	90	90	90	100	85	91.66	98.33	83.33	86.66	100

Table 2. Speaker type classification accuracy for the third and fourth experiment. Full set refers to the full hVd signals while segmented refers to the steady state portion of the vowel extracted from the full hVd signal.

		Classification results											
		Full set 96 kHz				Full set 48 kHz				Segmented set 48 kHz			
Experiment	Vowels	Child	Female	Male	Cumulative	Child	Female	Male	Cumulative	Child	Female	Male	Cumulative
Third	Combined	90	93.33	91.66	91.66	83.3	91.66	90	90	83.33	86.66	86.66	85.55
Fourth	had	100	100	100	100	100	100	100	100	100	100	100	100
	heed	100	100	100	100	100	100	100	100	100	100	100	100
	herd	100	100	100	100	100	100	100	100	100	100	100	100
	hawd	100	100	100	100	100	100	100	100	100	100	100	100
	hayed	100	100	100	100	100	100	100	100	100	100	100	100
	who'd	90	100	100	96.66	80	100	100	93.33	80	100	100	93.33

Taken together, the current data highlight the utility of information from the high-frequency spectrum of the speech signal (information beyond the traditionally studied low-frequency region) from male, female, and child speakers for automated speech and speaker recognition purposes. As such, the potential contributions of high-frequency energy (above 3-4 kHz) in naturally produced vowels produced by various speaker types, including children, should be the subject of further study. Additionally, research examining the use of both low- and high-frequency information for classification purposes may provide useful information regarding the potential benefits of concurrently using data from both spectral regions. Given the use of two speakers from each category in the current study, future research should investigate classification performance from a larger dataset of speakers, including the evaluation of algorithm performance on previously unseen data (*i.e.*, speech signals from novel speakers). This is expected to provide important information regarding how well the algorithm described in the current experiments recognizes novel speech data.

References

1. Molis MR, Diedesch, A Gallun F, Leek MR. Vowel identification by amplitude and phase contrast. J Assoc Res Otolaryngol2013;14:125-137.
2. Gelfer MP, Mikos VA. The relative contributions of speaking fundamental frequency and formant frequencies to gender identification based on isolated vowels. J Voice 2005;19:544-54.
3. Donai JJ, Paschall DD. Identification of high-pass filtered male, female, and child vowels: The use of high-frequency cues. J Acous Soc Am 2015;137:1971-82.
4. Vitela AD, Monson BB, Lotto AJ. Phoneme categorization relying solely on high-frequency energy. J Acous Soc Am 2014;137:65-70.
5. Monson BB, Lotto AJ, Story BH. Gender and vocal production mode discrimination using the high frequencies for speech and singing. Fron Psychol 2014;5:1-7.
6. Donai JJ, Lass NJ. Gender identification from high-pass filtered vowel segments; the use of high-frequency energy. Att Percep Psychophy 2015;77:2452-62.
7. Mishra N, Shrawankar U, Thakare VM. Automatic speech recognition using template model for man-machine interface. Proc. Int. Conf. ICAET, Chennai, India; 2010.
8. Vishwanathan S, Smola A, Vidal R. Binet-cauchy kernels on dynamical systems and its application to the analysis of dynamic scenes. Int J Comput Vision 2007;73:95-119.
9. Chaudhry R, Ravichandran A, Hager G, Vidal R. Histograms of oriented optical flow and binet-cauchy kernels on nonlinear dynamical systems for the recognition of human actions. Proc Cvpr IEEE 2009;5-7:1932-9.
10. Schölkopf B, Smola AJ. Learning with Kernels. Cambridge, MA: MIT Press; 2001.
11. Davis S, Mermelstein P. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. IEEE Trans Acoust Speech Signal Process 1980;28:357-66.
12. Kinnunen T, Karkkainen I, Franti P. Is speech data clustered? - Statistical analysis of cepstral features. Proc. 7th European Conference on Speech Communication and Technology (Eurospeech 2001), Aalborg, Denmark, 2001;4:2627-30.
13. Hayakawa S, Itakura F. The influence of noise on the speaker recognition performance using the higher frequency band. ICAS-SP-95 Acoust Speech Signal Process 1995;1:321-4.
14. Hu G, Wang D. Monaural speech segregation based on pitch tracking and amplitude modulation. IEEE Trans Neural Net 2004;15:1135-50.
15. Deshpande MS, Holambe RS. Robust speaker identification in the presence of car noise. Int J Biom 2011;3:189-205.