PAPER

Homomorphic Filtered Spectral Peaks Energy for Automatic Detection of Vowel Onset Point in Continuous Speech

Xian ZANG[†], *Nonmember and* **Kil To CHONG**[†a)], *Member*

SUMMARY During the production of speech signals, the vowel onset point is an important event containing important information for many speech processing tasks, such as consonant-vowel unit recognition and speech end-points detection. In order to realize accurate automatic detection of vowel onset points, this paper proposes a reliable method using the energy characteristics of homomorphic filtered spectral peaks. The homomorphic filtering helps to separate the slowly varying vocal tract system characteristics from the rapidly fluctuating excitation characteristics in the cepstral domain. The distinct vocal tract shape related to vowels is obtained and the peaks in the estimated vocal tract spectrum provide accurate and stable information for VOP detection. Performance of the proposed method is compared with the existing method which uses the combination of evidence from the excitation source, spectral peaks, and modulation spectrum energies. The detection rate with different time resolutions, together with the missing rate and spurious rate, are used for comprehensive evaluation of the performance on continuous speech taken from the TIMIT database. The detection accuracy of the proposed method is 74.14% for ±10 ms resolution and it increases to 96.33% for ±40 ms resolution with 3.67% missing error and 4.14% spurious error, much better than the results obtained by the combined approach at each specified time resolution, especially the higher resolutions of ±10 ∼ ±30 ms. In the cases of speech corrupted by white noise, pink noise and f-16 noise, the proposed method also shows significant improvement in the performance compared with the existing method.

key words: *vowel onset point, homomorphic filtering, peaks energy, vocal tract spectrum, noise robustness*

1. Introduction

Speech signals are produced by exciting the time-varying vocal tract (VT) system with an excitation sequence. The changes taking place in the speech production system are manifested as events in the speech signal. Motivated by this nature, if certain events of significance can be identified and detected, the analysis for feature extraction can be anchored around such events. This kind of event-based analysis is likely to produce consistent representation of speech information for applications in speech signal processing [1].

The vowel onset point (VOP) is one such important event at which there is a significant change in both VT and excitation source. VOP is defined as the instant that the onset of vowels takes place [1]–[3], or termed as the segmenting point in consonant-vowel (CV) transitions [4], [5]. Some earlier investigations [6]–[8] show that the very important and discriminatory information for speech analysis lies in a small region round the VOPs in each CV unit. Since the syllable-like CV units are basic units appropriate to describe the production and recognition aspects of speech [9], the VOPs can be used as anchor points to extract information for many speech processing tasks, such as: speech recognition [3], [10], speaker recognition [11], speech segmentation [12], [13], end-points detection [14], [15], keywords spotting [16], and positioning of pitch movement [17].

To realize good performance of the VOP event-based analysis, the first essential step is to develop a reliable method for automatic detection of VOPs. This issue did not attract specific attention until the 1990s. Hermes [2] first proposed a method for automatic VOP detection in fluent speech by identifying the points at which there is a rapid increase in the vowel strength. Other methods were later developed, such as the use of temporal information on intensity and rough spectral envelope [18]; energy profile, zero-crossing rate, and pitch information [19]; energy derivative [20]; neural network [12], [21]–[23]; wavelet transform [4]; phone model [24]; and reassignment spectra [25]. All of these methods mainly use VT system features in one form or another, but the excitation features are ignored. Thus, more recently, some works have focused on exploring the excitation source information for detecting VOPs. Prasanna et al. [26] proposed a combined method for VOP detection using excitation source energy, spectral peaks energy and modulation spectrum energy. Compared with the best performing individual VOP detection method, the combined method demonstrates a reduced total error rate and high detection rate within ±40 ms, but the performance degrades drastically at higher time resolution. This is mainly attributed to the poor performance of each individual method at the corresponding time resolutions. To increase the time resolution, Vuppala et al. [27] proposed a second-level procedure using uniform epoch intervals present in the vowel region to correct the positions of the hypothesized VOPs from combined evidence. However, this approach increases time cost.

Addressing this issue from the viewpoint of efficiency and low complexity, we propose an improved feature in this paper, named homomorphic filtered spectral peaks energy. As mentioned in the beginning, speech signals are produced by exciting the time-varying VT system with excitation sequence. They are represented as two multiplied components in the discrete Fourier transform (DFT) spectrum. The VT system features used in previous research are mainly based on the DFT spectrum, thus the mixed information reduce

Manuscript received October 29, 2012.
Manuscript revised December 17, 2012.

[†]The authors are with the Department of Electronic Engineering, Jeonbuk National University, Jeonju-si, Jeollabuk-do, 561–756, R.O. Korea.
a) E-mail: kitchong@jbnu.ac.kr (Corresponding author)

DOI: 10.1587/transinf.E96.D.949

the purity of features and consequently degrade the performance. To solve the problem, we adopt homomorphic filtering to separate the slowly varying vocal tract system characteristics from the rapidly fluctuating excitation characteristics in the cepstral domain. In the VT spectrum reconstructed based on the separated VT components, the filtered peaks indicating the VT shape related to vowels are distinct, and their amplitudes are extracted to form the time-varying VT energy. We exploit the significant changes in the energy level to realize VOP detection.

The paper is organized as follows: Sect. 2 explains the homomorphic speech analysis for transforming the multiplied components of VT system and excitation source in the DFT spectrum to linear combined form in the cepstrum and proposes the energy characteristics of the homomorphic filtered spectral peaks for VOP detection. Section 3 presents four experiments to analyze the performance of the proposed method under the cases of clean and noise-corrupted speech, gender-dependent speech, and various difficult continuous utterances. The results are also discussed. Section 4 provides conclusions and the direction of future work.

2. VOP Detection Based on Homomorphic Filtered Spectral Peaks Energy

In the process of pronouncing vowels, the vocal tract acts as a resonant cavity. In this situation, the distinct vocal tract shape is represented by prominent peaks, called formants, in the spectrum. This unique character inspires that the large energy of these formants could be used as feature to detect VOP. However, the DFT spectrum consists of the multiplication of excitation and vocal tract system components. The formants cannot be easily identified since they are smeared in the harmonics resulting from the excitation source, which introduce redundant speaker-dependent information. If these two kinds of components could be separated from each other, the clear formants structure could be obtained to provide accurate and stable information for VOP detection.

However, it is difficult to separate them in the frequency domain because of the relationship of multiplication between them. In order to solve the problem, we adopt homomorphic filtering to facilitate the separation in cepstral domain, where the linear combination of the two components is realized. Based on the separated VT system components, the VT spectrum is reconstructed, where the spectral peaks are free from excitation information. Thus the energy of these spectral peaks could be extracted as a pure vocal tract system feature for VOP detection.

Consider a frame of time-domain speech signal (Fig. 1 (a)), we construct its spectrum via a K-point DFT operation, expressed as

$$X[n,k] = \frac{1}{M}\sum_{m=0}^{M-1} x(n,m)\, e^{-j2\pi k\frac{m}{M}}, 0 \le k \le K-1 \qquad (1)$$

where n is the frame index, M is the number of sampling points in each frame, and k is the DFT coefficient index. The corresponding magnitude spectrum $|X[n,k]|$ is shown in Fig. 1 (b), where each peak contains mixed information of VT system and excitation. Next, we carry on homomorphic filtering to retain the VT system components merely.

The real cepstrum $C[n,l]$ (Fig. 1 (c)) of the DFT magnitude spectrum is defined as

$$C[n,l] = \frac{1}{K}\sum_{k=0}^{K-1} \log|X[n,k]|\, e^{j2\pi l\frac{k}{K}}, 0 \le l \le K-1 \qquad (2)$$

In the cepstrum, the components in the lower quefrency part are representative of the slowly varying VT characteristics in the spectrum, while the high quefrency components correspond to the more rapidly fluctuating excitation characteristics. In this way, the multiplied two components are split up into two linear combined parts. Thus we could use a low-time liftering window to extract the cepstral components related to VT characteristics. The low-time liftering window is given by [28]

$$w[l] = \begin{cases} 1, & l \le L_1 \\ 0.5(1+\cos[\pi(l-L_1)/\Delta L]), & L_1 < l \le L_1+\Delta L \\ 0, & l > L_1+\Delta L \end{cases} \qquad (3)$$

where $L = L_1 + \Delta L$ is the cut-off length of the liftering window, and is typically 15 or 20. The section ΔL is tapered, and its length can be as much as 25% of L. Multiplying the whole cepstrum by the low-time liftering window, we extract the cepstral VT characteristics, indicated in Eq. (4) and shown in Fig. 1 (d),

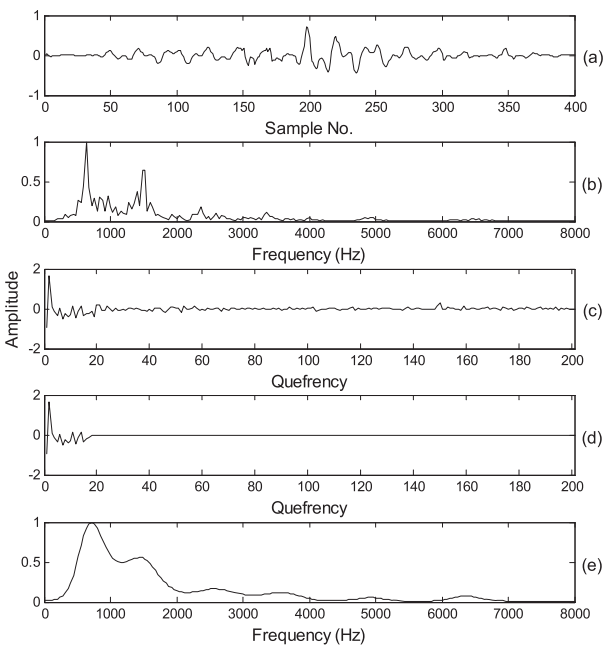

Fig. 1 (a) Time-domain waveform of a vowel frame; (b) DFT magnitude spectrum; (c) Real cepstrum; (d) VT characteristics in cepstrum by low-time liftering; (e) Estimated VT spectrum of the vowel frame.

$$C_h[n,l] = w[l]\,C[n,l], \quad 0 \le l \le K-1 \tag{4}$$

Applying the inverse transformation to the low-time liftered sequence yields the homomorphic filtered smooth spectrum, shown in Fig. 1 (e), which is the VT spectrum of the given short term speech. Compared with the original DFT magnitude spectrum, the peaks indicating formants become more distinct since the excitation source information is eliminated. Due to the fact that the spectrum amplitude decreases with increasing frequency, formants above 5500 Hz have low amplitudes with no considerable difference from those in cases of non-vowel frames. Therefore, they are not significant in forming substantial changes in the energy level for VOP detection. In contrast, the five largest formants below 5500 Hz emphasize the distinct energy feature of vowels. On the other hand, there exist influencing consonants such as semivowels which share similar characteristics of vowels; normally possessing three distinguishable formants in contrast to the five in vowels. In this manner, we consider two more useful peaks to enhance the dissimilarity between the energies of these consonants and vowels. Given the above-mentioned reasons, we select the five largest homomorphic filtered spectral peaks from the VT spectrum and sum their amplitudes to be used as the energy feature for each frame.

Figure 2 (a) shows the continuous speech of *Don't ask me to carry an oily rag like that*, taken from the TIMIT database, where the circles denote the reference VOPs. The speech is divided into frames of 25 ms (sampling frequency is 16000 Hz) by a Hamming window with a shift of 10 ms. For each frame, a 400-point DFT is computed, and the VT spectrum is then separated from the DFT spectrum by the homomorphic filtering detailed above. We choose the five largest peaks from each VT spectrum and plot the sum of their amplitudes as a function of time to represent the time-varying VT energy, as shown in Fig. 2 (b). Since the energy will reach peak when pronouncing vowel, the onset of a vowel can be observed as significant change in the sum of five largest peaks in the VT spectrum. To enhance the change at the VOP, some unnecessary peaks are eliminated by two steps [26]. The first one is computing the sum of slops in a short duration centered at each peak with the help of first-order difference (FOD), and the peaks with sum values lower than half the mean value are eliminated. The second one is based on the assumption that two VOPs rarely occur within a 50-ms interval; hence if two adjacent peaks happen to be within 50 ms, then the lower peak is eliminated. After this enhancement, the reserved peaks in the sum of five largest peaks in the VT spectrum are represented by asterisks in Fig. 2 (b). With respect to each of these local peaks, the nearest local minimum on either side is identified and marked with circles in Fig. 2 (b). The amplitudes in each segment bounded by two successive local minima are then normalized to [0, 1], as shown in Fig. 2 (c). From this normalized version of the time-varying VT energy, the sharply rising instants related to the VOPs can be easily observed. To automatically detect such instants, a modulated Gaussian window (MGW) [29] of length 100 ms is used with a suitable shape shown in Fig. 3. Figure 2 (d) shows the result of convoluting the normalized values with the MGW, which is called the VOP candidates plot using homomorphic filtered spectral peaks energy. The peaks valid for VOPs are selected from the candidates using a peak-picking algorithm based on the logic that there should be a negative region between two successive peaks, since the negative region is a symbol of transition to the next vowel region. Otherwise, only the larger peak is valid for VOP. For instance, in Fig. 2 (d), there are two peaks around $t = 0.3s$ with no negative region between them, which means they are in the same vowel structure. Since the first peak is higher, indicating a sharply rising point, it is hypothesized as the VOP, and the second one is taken as an unwanted peak. The sentence has 12 reference VOPs, which are marked with circles in Fig. 2 (a). The proposed method hypothesizes 12 VOPs marked with asterisks in Fig. 2 (d), and all of them are matched with the cor-

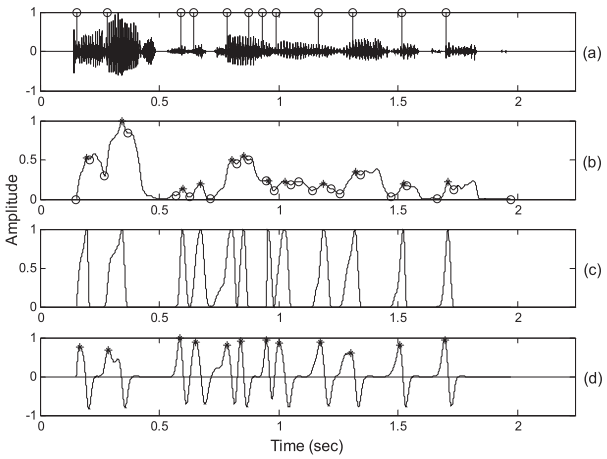

Fig. 2 (a) Time-domain waveform of the speech of *Don't ask me to carry an oily rag like that*; (b) Sum of the five largest peaks in the VT spectrum; (c) Enhanced version of (b); (d) VOP candidates plot using homomorphic filtered spectral peaks energy.

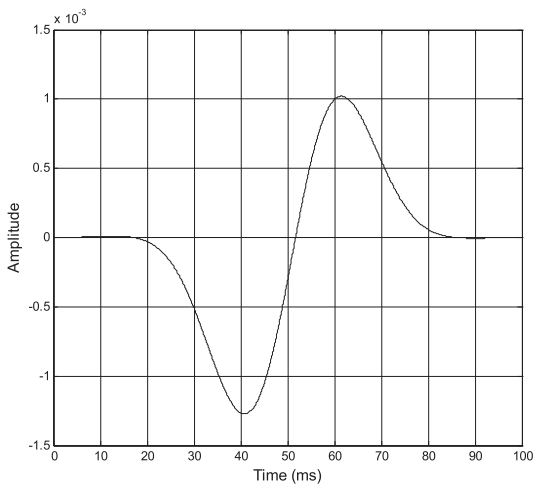

Fig. 3 Modulated Gaussian Window.

responding reference VOPs within small deviation, respectively. In other word, all the reference VOPs are detected successfully.

3. Experiments and Discussions

VOP detection experiments are conducted on the TIMIT speech corpus. We choose the test set consists of 168 speakers with 56 females and 112 males grouped into eight different dialect regions. The reference VOPs of each sentence are manually marked according to the associated phonetic transcription files.

To evaluate the performance of the proposed method, we adopt three criteria as follows:

1) Detection rate with a resolution of T (DETR_T), the percentage of reference VOPs matched by hypothesized VOPs within $\pm T$ ms deviation with respect to the total of reference VOPs;
2) Missing rate (MISSR), the percentage of undetected reference VOPs with respect to the total of reference VOPs after the maximum resolution of $\pm T$ ms;
3) Spurious rate (SPUR), the percentage of unmatched hypothesized VOPs with respect to the total of reference VOPs after the maximum resolution of $\pm T$ ms.

Here, the three criteria are proposed under the consideration of comprehensive error analysis. On the one hand, we desire high detection rate; on the other hand, we also wish the number of unmatched hypothesized VOPs as few as possible. Hence, if the method realizes higher DETR_T with lower MISSR and SPUR, the performance is better.

In the first experiment, performance of the proposed method on clean speech is compared with Prasanna's [26] results of VOP detection using excitation source (EXC), spectral peaks (VT), modulation spectrum (MOD) and combined (COMB) method. The evaluation is based on two sentences *Don't ask me to carry an oily rag like that*, and *She had your dark suit in greasy wash water all year* selected from each of the 168 speakers, There are 25 VOPs in these two sentences per person and hence a total of 4200 VOPs.

The performance of each different method for clean utterances is given in Table 1. Column 1 indicates different methods considered in the analysis for VOP detection. Column 2 indicates the total number of VOPs hypothesized by each method. Columns 3-6 indicate the percentage of VOPs detected within specified time resolution. Columns 7 and 8 indicate the percentage of missed reference VOPs and spurious hypothesized VOPs, respectively. From the results, it is evident that the performance of the combined method is better than the individual methods (EXC, VT and MOD). However, the detection accuracy decreases significantly at higher time resolutions ($\pm 10 \sim \pm 30$ ms). The proposed method is observed to be superior to the combined method. By comparison, about 20% more VOPs are detected within ± 10, ± 20, and ± 30 ms deviations; besides, both the missing error and spurious error are maintained at a low level. This demonstrates that the homomorphic filtered spectral peaks carry more accurate VOP information. Although the performance of the proposed method at ± 10 ms is limited due to the high similarity of the signal characteristics preceding and following the VOP when voiced consonants present before vowel, the high detection rate at resolutions of $\pm 20 \sim \pm 40$ ms may be sufficient for applications such as end-point detection and identification of voiced/unvoiced regions.

In the second experiment, to verify the robustness of the features to noise, the same clean test utterances consists of 4200 VOPs are corrupted using white noise, pink noise and f-16 noise from the NOISEX92 database, at SNRs of 20, 10, and 0 dB. The results of VOP detection using the aforementioned methods under three different noises environment at various SNR levels are given in Tables 2–4, respectively. Compared with the detection rate for the clean speech, the performance of the individual methods (EXC, VT and MOD), especially MOD, degrades significantly in each noisy case. For this reason, the improvement in the performance of combined method suffers from a great limitation. In the cases of 10 and 20 dB SNR, the proposed method maintains a relatively higher detection rate from about 85% at ± 40 ms resolution to about 65% at ± 10 ms resolution, much greater than the results obtained by the best performing method among EXC, VT, MOD, and COMB at each corresponding resolution. Moreover, the missing and spurious error of the proposed method are both lower than 17%, but they are more than 30% in the combined approach. In the noisy case of 0 dB, although the proposed method has almost same level in the missing and spurious error as combined method, it ensures the detection rate more than 45% at the highest time resolution. By this comparison, it is obvious that the homomorphic filtered spectral peaks energy

Table 1 Performance of VOP detection using excitation source (EXC), spectral peaks (VT), modulation spectrum (MOD), combined (COMB) and proposed methods for clean TIMIT database consists of 4200 VOPs.

Method	Number of Hypothesized VOPs	DETR_T (% within ms)				MISSR (%)	SPUR (%)
		±10	±20	±30	±40		
EXC	4219	37.40	51.20	63.20	95.60	4.40	4.86
VT	4256	29.30	57.20	76.60	94.43	4.57	5.90
MOD	3977	38.30	47.70	69.30	92.62	7.38	2.07
COMB	4170	52.10	63.10	73.10	96.02	3.98	3.26
Proposed	4220	74.14	87.50	93.12	96.33	3.67	4.14

Table 2 Performance of VOP detection using excitation source (EXC), spectral peaks (VT), modulation spectrum (MOD), combined (COMB) and proposed methods for TIMIT database consists of 4200 VOPs under white noise environment at various SNR levels.

SNR (dB)	Method	Number of Hypothesized VOPs	DETR_T (% within ms)				MISSR (%)	SPUR (%)
			±10	±20	±30	±40		
20	EXC	4237	19.36	37.31	54.17	65.02	34.98	35.86
	VT	4310	27.12	46.07	54.90	65.29	34.71	37.33
	MOD	4013	16.48	30.00	41.02	51.24	48.76	44.31
	COMB	4238	17.95	38.69	56.60	68.45	31.55	32.45
	Proposed	4224	68.07	76.74	82.79	86.36	13.64	14.21
10	EXC	4454	22.12	42.31	51.45	63.12	36.88	42.93
	VT	4458	24.31	40.29	50.10	56.93	43.07	49.21
	MOD	4041	17.69	26.67	37.43	43.86	56.15	52.36
	COMB	4260	19.05	39.38	56.14	68.81	31.19	32.62
	Proposed	4233	64.62	74.31	81.50	85.07	14.93	15.71
0	EXC	4471	17.69	31.02	47.43	57.43	42.57	49.02
	VT	4569	12.57	27.69	41.29	50.26	49.74	58.52
	MOD	4228	13.88	25.55	35.43	43.40	56.60	57.26
	COMB	4292	19.50	40.12	58.93	67.29	32.71	34.90
	Proposed	4302	45.83	56.88	64.24	69.79	30.21	32.64

Table 3 Performance of VOP detection using excitation source (EXC), spectral peaks (VT), modulation spectrum (MOD), combined (COMB) and proposed methods for TIMIT database consists of 4200 VOPs under pink noise environment at various SNR levels.

SNR (dB)	Method	Number of Hypothesized VOPs	DETR_T (% within ms)				MISSR (%)	SPUR (%)
			±10	±20	±30	±40		
20	EXC	4278	20.69	39.79	55.76	68.55	31.45	33.31
	VT	4318	25.31	44.00	58.10	67.12	32.88	35.69
	MOD	4005	16.43	30.64	39.24	49.21	50.79	46.14
	COMB	4171	19.31	40.02	57.38	68.64	31.36	30.67
	Proposed	4199	69.31	77.98	83.93	86.86	13.14	13.12
10	EXC	4298	19.24	39.74	55.62	67.12	32.88	35.21
	VT	4337	22.07	38.88	51.79	60.52	39.48	42.74
	MOD	4016	14.98	26.29	35.86	44.83	55.17	50.79
	COMB	4186	18.76	38.60	55.50	67.12	32.88	32.55
	Proposed	4282	67.21	76.05	81.45	85.62	14.38	16.33
0	EXC	4417	18.43	38.88	54.86	63.90	36.10	41.26
	VT	4409	18.12	31.50	41.52	49.07	50.93	55.90
	MOD	4276	14.79	26.29	35.86	44.55	55.45	57.26
	COMB	4283	18.69	37.95	54.31	66.31	33.69	35.67
	Proposed	4329	47.57	58.05	65.76	71.93	28.07	31.14

feature yields much better robustness to noises than the combination of individual evidence from EXC, VT, and MOD.

The next experiment conducted is to analyze the gender-dependent performance of the proposed VOP detection method. The database used earlier is considered in this case also. The 168 speakers are composed of 112 male speakers with 2800 VOPs and 56 female speakers with 1400 VOPs. The whole database is classified into two groups according to their gender. The performance of the proposed method for each group is shown in the Table 5. In the male case, the detection accuracy is 74.00% for ±10 ms resolution and increases to 96.57% for ±40 ms resolution. In the female case, the detection accuracy is 74.43% for ±10 ms resolution and increases to 95.86% for ±40 ms resolution. It is observed that there is only slight difference in the detection rate, missing error and spurious error between the two groups. The nearly same performance demonstrates that the gender-independency of the proposed method.

Finally, to analyze the performance of the proposed VOP detection method with respect to different continuous speech, another 100 different utterances are selected form the test set. In this experiment, we focus on some difficult cases of continuous speech, such as the sentence of *Where were you while we were away* with many semivowels highly articulated with vowels and the sentence of *Thus technical efficiency is achieved at the expense of actual experience* with polysyllabic words containing multiple vowels. There are a total of 926 VOPs under this situation in the 100 sen-

Table 4 Performance of VOP detection using excitation source (EXC), spectral peaks (VT), modulation spectrum (MOD), combined (COMB) and proposed methods for TIMIT database consists of 4200 VOPs under f-16 cockpit noise environment at various SNR levels.

SNR (dB)	Method	Number of Hypothesized VOPs	DETR_T (% within ms)				MISSR (%)	SPUR (%)
			±10	±20	±30	±40		
20	EXC	4418	19.26	38.29	55.62	67.71	32.29	37.48
	VT	4336	19.83	36.76	53.21	65.00	35.00	38.24
	MOD	4016	14.79	27.62	39.38	49.76	50.24	45.86
	COMB	4356	17.90	37.76	55.86	67.86	32.14	35.86
	Proposed	4249	67.71	76.33	82.33	85.76	14.24	15.40
10	EXC	4472	18.62	37.33	53.12	65.12	34.88	41.36
	VT	4350	20.64	36.55	56.10	65.81	34.19	37.76
	MOD	4021	14.19	25.45	35.21	43.86	56.14	51.88
	COMB	4334	18.00	38.50	53.67	64.93	35.07	38.26
	Proposed	4300	64.83	74.64	81.81	85.67	14.33	16.71
0	EXC	4492	19.43	36.90	51.12	61.43	38.57	45.52
	VT	4455	13.62	27.69	42.52	54.02	45.98	52.05
	MOD	4278	14.57	26.19	35.57	44.14	55.86	57.71
	COMB	4376	17.95	36.86	52.12	65.71	34.29	38.48
	Proposed	4498	49.60	62.05	69.26	75.19	24.81	31.90

Table 5 Performance of the proposed VOP detection method for 112 male speakers with 2800 VOPs and 56 female speakers with 1400 VOPs.

Gender	Number of Reference VOPs	Number of Hypothesized VOPs	DETR_T (% within ms)				MISSR (%)	SPUR (%)
			±10	±20	±30	±40		
Male	2800	2839	74.00	87.79	93.54	96.57	3.43	4.82
Female	1400	1381	74.43	86.93	92.29	95.86	4.14	2.79

Table 6 Performance of the proposed VOP detection method for 100 difficult continuous speech with 926 VOPs.

Number of Reference VOPs	Number of Hypothesized VOPs	DETR_T (% within ms)				MISSR (%)	SPUR (%)
		±10	±20	±30	±40		
926	880	52.70	68.57	79.05	86.61	13.39	8.42

tences. The performance of the proposed method is shown in Table 6. It is observed that the number of hypothesized VOPs is less than the number of reference VOPs. This may be mainly due to the closely ranked vowels in the polysyllabic word; the transition from a vowel to another vowel is not distinct enough for detection, thus some VOPs are missed. Compared with the earlier cases of two sentences, the performance in detection rate at each specified resolution is relative poorer. This may be attributed to the presence of highly confusable voiced consonants units, such as semivowels. Since their characteristics are very similar with vowels, the sharply rising instants related to the VOPs in the time-varying homomorphic filtered spectral peaks energy may be located in a larger deviation, accordingly the detection accuracy is reduced. The detection rate of 86.61% within ±40 ms is acceptable but it is not satisfying at higher time resolution, the present method needs to be improved to solve the problem in this situation.

4. Conclusions

In this paper, an improved method for VOP detection with low complexity is proposed. The method explores the features of VT system represented by homomorphic filtered spectral peaks energy. The homomorphic filtering is used to transform the multiplied components of VT and excitation in the DFT spectrum to linear combined form in the cepstral domain and then realize the separation of them. The VT spectrum reconstructed based on the separated components possesses the distinct peaks indicating the VT shape related to vowels, which are free from the redundant information existing in excitation source. The amplitudes of the homomorphic filtered spectral peaks are extracted as energy feature and the sharply rising instants in the energy level are exploited for detecting VOPs.

The proposed method is compared with the existing approach using the combination (COMB) of evidence from the

excitation source (EXC), spectral peaks (VT), and modulation spectrum (MOD) energies. Since the combined method needs three kinds of features to realize enhancement of VOP information, the computation complexity is higher than the proposed method. Moreover, in both the cases of clean and noise-corrupted speech, the proposed method shows significant improvement in the performance compared to the combined approach, especially at the higher resolutions, about 20% more VOPs are detected. Further improvement should focus on the performance at a time resolution of ±10 ms and robustness in difficult cases of speech to satisfy the needs of different practical applications in speech processing.

Acknowledgments

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korean government (MEST) (No. 2011-0027689).

References

[1] S.R. M. Prasanna, S.V. Gangashetty, and B. Yegnanarayana, "Significance of vowel onset point for speech analysis," Proc. 6th Biennial Conf. Signal Process. Commun., IISc – Bangalore, India, July 2001.

[2] D.J. Hermes, "Vowel onset detection," J. Acoust. Soc. Am., vol.87, no.2, pp.866–873, Feb. 1990.

[3] C.C. Sekhar, J.Y.S.R.K. Rao, and B. Yegnanarayana, "Recognition of consonant-vowel (CV) units of speech in Indian languages," Proc. National Seminar Inf. Revolution and Indian Lang., Hyderabab, pp.22.1–22.6, Nov. 1999.

[4] M.F. Tolba, T. Nazmy, A.A. Abdelhamid, and M.E. Gadallah, "A novel method for Arabic consonant/vowel segmentation using wavelet transform," Int. J. Intell. Coop. Inf. Syst., vol.5, no.1, pp.353–364, July 2005.

[5] M.T. Lin, C.K. Lee, and C.Y. Lin, "Consonant/vowel segmentation for mandarin syllable recognition," Comput. Speech Lang., vol.13, no.3, pp.207–222, July 1999.

[6] D. Kewley-Port, D.B. Pisoni, and M. Studdert-Kennedy, "Perception of static and dynamic acoustic cues to place of articulation in initial stop consonants," J. Acoustic. Soc. Am., vol.73, no.5, pp.1779–1793, May 1983.

[7] M.E. Tekieli and W.L. Cullinan, "The perception of temporally segmented vowels and consonant-vowel syllables," J. Speech Hear. Res., vol.22, pp.103–121, 1979.

[8] S. Furui, "On the role of spectral transition for speech perception," J. Acoust. Soc. Am., vol.80, no.4, pp.1016–1025, 1986.

[9] S. Greenberg, "Speaking in shorthand—a syllable-centric perspective for understanding pronunciation variation," Speech Commun., vol.29, no.2, pp.159–176, Nov. 1999.

[10] C.C. Sekhar and B. Yegnanarayana, "A constraint satisfaction model for recognition of stop consonant–vowel (SCV) utterances," IEEE Trans. Speech Audio Process., vol.10, no.7, pp.472–480, Oct. 2002.

[11] L. Mary and B. Yegnanarayana, "Extraction and representation of prosodic features for language and speaker recognition," Speech Commun., vol.50, no.10, pp.782–796, Oct. 2008.

[12] C.C. Sekhar and B. Yegnanarayana, "Neural netwok models for spotting stop consonant-vowel (SCV) segments in continuous speech," Proc. IEEE Int. Conf. Neural Networks, vol.4, pp.2003–2008, Washington, D.C, USA, June 1996.

[13] H. Kasuya and H. Wakita, "An approach to segmenting speech into vowel- and nonvowel-like intervals," IEEE Trans. Acoust. Speech, Signal Process., vol.ASSP-27, no.4, pp.319–327, Aug. 1979.

[14] S.R.M. Prasanna, J.M. Zachariah, and B. Yegnanarayana, "Begin-end detection using vowel onset points," Proc. Workshop Spoken Lang. Process., TIFR Mumbai, India, Jan. 2003.

[15] B. Yegnanarayana, S.R.M. Prasanna, and C.S. Gupta, "Combining evidence from source, suprasegmental and spectral features for a fixed-text speaker verification system," IEEE Trans. Speech Audio Process., vol.13, no.4, pp.575–582, July 2005.

[16] B.V.S. Reddy, K.V. Rao, and S.R.M. Prasanna, "Keyword spotting using vowel onset point, vector quantization and hidden Markov modeling based techniques," Proc. IEEE Region 10 Conf., Hyderabad, India, Nov. 2008.

[17] D.J. Hermes, "Timing of pitch movement and accentuation of syllables," Proc. Int. Conf. Spoken Lang. Process., Philadelphia, PA, USA, Oct. 1996.

[18] R.W.L. Kortekaas, D.J. Hermes, and G.F. Meyer, "Vowel-onset detection by vowel-strength measurement, cochlear-nucleus simulation, and multiplayer perceptrons," J. Acoust. Soc. Am., vol.99, no.2, pp.1185–1199, Feb. 1996.

[19] J.F. Wang, C.H. Hu, S.H. Chang, and J.Y. Lee, "A hierarchical neural network model based C/V segmentation algorithm for isolated mandarin speech recognition," IEEE Trans. Signal Process., vol.39, no.9, 2141–2146, Sept. 1991.

[20] C.C. Sekhar, Neural network models for recognition of stop consonant-vowel (SCV) segments in continuous speech, Ph.D. thesis, Dept. of Comput. Sci. and Eng., Indian Inst. of Technol. Madras, Chennai, India, April 1996.

[21] J.Y.S.R.K. Rao, C.C. Sekhar, and B. Yegnanarayana, "Neural networks based approach for detection of vowel onset point," Proc. Int. Conf. Adv. in Pattern Recogn. Digit. Tech., Calcutta, India, pp.316–320, Dec. 1999.

[22] S.V. Gangashetty, C.C. Sekhar, and B. Yegnanarayana, "Detection of vowel onset points in continuous speech using autoassociative neural network models," Proc. Int. Conf. Spoken Lang. Process., pp.1081–1084, Jeju Island, Korea, Oct. 2004.

[23] S.V. Gangashetty, C.C. Sekhar, and B. Yegnanarayana, "Spotting multilingual consonant-vowel units of speech using neural network models," Nonlinear Analyses and Algorithms for Speech Process., vol.3817, pp.303–317, 2006.

[24] A. Kazemzadeh, J. Tepperman, J. Silva, H. You, S. Lee, A. Alwan, and S. Narayanan, "Automatic detection of voice onset time contrasts for use in pronunciation assessment," Proc. Int. Conf. Spoken Lang. Process., pp.721–724, Pittsburgh, PA, USA, Sept. 2006.

[25] V. Stouten and H.V. Hamme, "Automatic voice onset time estimation from reassignment spectra," Speech Commun., vol.51, no.12, pp.1194–1205, Dec. 2009.

[26] S.R.M. Prasanna, B.V.S. Reddy, and P. Krishnamoorthy, "Vowel onset point detection using source, spectral peaks, and modulation spectrum energies," IEEE Trans. Audio Speech Lang. Process., vol.17, no.4, pp.556–565, May 2009.

[27] A.K. Vuppala, K.S. Rao, and A. Chakrabarti, "Improved vowel onset point detection using epoch intervals," Int. J. Electron. Commun., vol.66, no.8, pp.697–700, Aug. 2012.

[28] R.W. Schafer and L.R. Rabiner, "System for automatic format analysis of voiced speech," J. Acoust. Soc. Am., vol.47, pp.634–648, Feb. 1970.

[29] S.R.M. Prasanna and B. Yegnanarayana, "Detection of vowel onset point events using excitation information," Proc. European Conf. Speech Commun. Technol., pp.1133–1136, Lisbon, Portugal, Sept. 2005.

Xian Zang received the MS degree in Electronic Engineering from Jeonbuk National University, Korea, in 2009. Currently, she is a Ph.D. candidate in the Department of Electronic Engineering, Jeonbuk National University, Korea. Her research interests are in the areas of speech signal processing.

Kil To Chong received the Ph.D. degree in Mechanical Engineering from Texas A&M University, U.S.A., in 1995. Currently, he is a professor in the Department of Electronic Engineering, Jeonbuk National University, Korea. His research interests are in the areas of motor fault detection, time-delay system, computer network technology, automatic control, marine navigation, robotics, neural networks, and artificial intelligence.