USE OF PITCH CONTINUITY FOR ROBUST SPEECH ACTIVITY DETECTION

Yiwen Shao[1,2] *and Qiguang Lin*[1]

[1]Baihu Technology Co., Ltd., China
[2]Center for Language and Speech Processing,
Johns Hopkins University, Baltimore, MD, 21218, USA
yshao18@jhu.edu, qlin@baihusoft.com

ABSTRACT

Speech activity detection (SAD) is an important component for various speech processing applications and has been researched extensively recently. The pitch continuity, a significant characteristic of speech, however, has not successfully played a role in existing SAD methods. In this work, we propose a novel way to integrate the pitch continuity with pitch-related features. Practice is carried out through the Combo-SAD approach: We examine three consecutive frames and assume that they all have the same pitch as the center frame due to pitch continuity. Corresponding feature values are recomputed at the adjusted pitch location and then used in the final expression. The new combo feature is evaluated with various types of additive noise at different signal-to-noise ratios (SNR). The results show that the new feature leads to better SAD performance (with an up to 39.3% relative improvement on miss rate compared to Combo-SAD). We also introduce a novel variant of the underlying autocorrelation function and illustrate how it can improve the accuracy of pitch detection.

Index Terms— autocorrelation function, speech activity detection, pitch continuity, pitch detection.

1. INTRODUCTION

Speech activity detection (SAD), i.e., discrimination of speech or nonspeech segments in an audio input, is a significant part in speech processing. Based on its outcome, much follow-up work can be done, such as speech coding, speech recognition, and speaker recognition. Recently, SAD has received increased interest due to the ARPAs RATS initiative and the NIST OpenSAD evaluation [1]. Many SAD methods/systems have been developed, along with much effort to compare their performance [2, 3].

SAD is relatively straightforward for clean speech inputs. What differentiates the various SAD methods is the robustness against noise interference. Energy based, or more sophisticated acoustic feature based SAD methods tend to produce more false alarms for noisy speech inputs. Phonation (or voicing) feature based methods, by applying the fact that all voiced sounds are periodic, are advantageous in mitigating noise that typically lacks the periodicity or has different periodicity than speech.

Among all the SAD methods based on phonation features, harmonics related features serve as a major ingredient, especially the pitch F0. For example, in the Combo-SAD technique [4], 3 of the 5 features depend on pitch. They show great robustness and deliver good SAD accuracy in various acoustic environments. However, these harmonics-dependent features heavily rely on the harmonics in individual frames, making it vulnerable to the interference of noise with strong harmonics. To solve it, we resort to another crucial property of speech — pitch continuity, that is, human speech exhibits a smooth, gradual pitch trajectory in speech. And we can reasonably assume that pitch is unchanged between the adjacent frames. But for noise segments, this phenomenon is not typically observed. Even though the noise may contain periodic or quasi-periodic signals, they usually come from multiple sources in the background and dont share a continuous pitch contour. For this reason, pitch continuity should be a powerful tool for us to distinguish noise from speech. Nonetheless, it has so far only been exploited in pitch detection algorithms [5], but has not been utilized in SAD. In this work, a novel method is proposed to incorporate pitch continuity into pitch-related features to improve the SAD performance.

Toward that goal, three features in the Combo-SAD approach [4] are selected as the working vehicle. The new features are all computed frame-by-frame, and the process can be described in three steps: 1) For each frame, locate the pitch F0; 2) Use it to compute the feature values of the current frame, as well as of two neighboring frames by assuming pitch continuity; and 3) Choose the median of them as the final feature value. The new method is evaluated with noisy speech data and shows a substantial improvement over the original one. Additionally, we introduce a modified form of the autocorrelation function for pitch detection. It is illustrated that the new form helps compensate for the windowing effects and prevents over-compensation from happening [6].

Work in part supported by Ministry of Science and Technology, China (Contract 14C-26213201061).

2. EXTENDED PITCH-BASED FEATURES

In this section, we first review the three pitch-related features in the Combo-SAD method [4]. Then we discuss the autocorrelation function (ACF) used in them. Based on the experiments on synthetic speech, we propose an improved variant of ACF for pitch detection. Finally, pitch continuity is integrated with these features to elevate the robustness of SAD.

2.1. Review of the Combo-SAD method

Sadjadi and Hansen [4] propose a 1-dimensional combo feature that is compressed from 5 features via Principal Component Analysis (PCA). Because our main purpose in this study is the use of pitch continuity, we focus on three of these five features, namely, harmonicity, clarity, and periodicity that are all related to the pitch. We leave out the prediction gain and perceptual spectral flux.

The procedure in which harmonicity, clarity, and periodicity are computed is briefly summarized below. Interested readers are referred to [4, 7] and the references therein for more details of the algorithms.

1) Harmonicity (or harmonics-to-noise ratio): the relative height of the maximum autocorrelation, $r_{xx,s}$ (we will discuss on it later), peak in the plausible pitch range (62.5 to 500 Hz, or the time domain equivalents of 16 ms to 2 ms):

$$H(t) = \frac{r_{xx,s}(t, k_{max})}{r_{xx,s}(t,0) - r_{xx,s}(t,k_{max})},$$
$$k_{max} = \underset{2ms \leq k \leq 16ms}{argmax} \;\; r_{xx,s}(t,k)$$

2) Clarity: the relative depth of the minimum average magnitude difference function (AMDF):

$$AMDF(t,q) \approx 0.8 \times \sqrt{2r_{xx,s}(t,0) - r_{xx,s}(t,q)},$$
$$C(t) = 1 - \frac{AMDF(t,q_{min})}{AMDF(t,q_{max})},$$
$$q_{min(max)} = \underset{2ms \leq q \leq 16ms}{argmin(max)} AMDF(t,q)$$

3) Periodicity: the maximum peak of the harmonic product spectrum (HPS) [9] in the short-time Fourier transform domain:

$$P(t) = HPS(t, \omega_{max}),$$
$$\omega_{max} = \underset{62.5Hz \leq \omega \leq 500Hz}{argmax} \;\; H(t,\omega)$$

where k_{max}, q_{min} and ω_{max} in Harmonicity, Clarity and Periodicity, respectively, are all different forms of the same entity, pitch. Both k_{max} and q_{min} are the reciprocal of F0, and ω_{max} is exactly the F0.

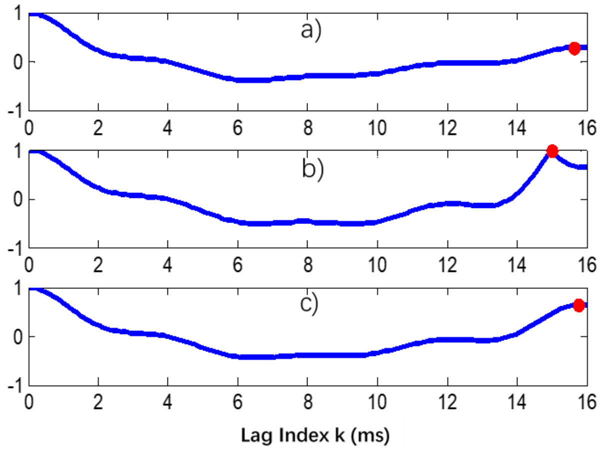


Fig. 1. Normalized ACFs: a) cACF; b) sACF; and c) eACF with $\beta = 0.5$. The signal has a fundamental period of 15.74 ms (or a pitch of 63.5 Hz). Solid dots denote k_{max} where the normalized ACF exhibits a peak.

2.2. Exponential ACF

Both harmonicity and clarity rely on a scaled autocorrelation function (ACF) as proposed by [6]. The scaled ACF (sACF) is defined by dividing the conventional ACF (cACF), $r_{xx,c}(t,k) = \sum_{j=0}^{N-1} x(j)w(j)x(j+k)w(j+k)$ by the ACF of the window function itself:

$$r_{xx,s}(t,k) = \frac{r_{xx,c}(t,k)}{\sum_{j=0}^{N-1} w(j)w(j+k)} \qquad (1)$$

where w(j) is a Hanning window of 32 ms long, x(j) is the input signal, and t and k are frame and autocorrelation lag indices, respectively. The frame shift rate is 10 ms. The purpose of the division is to compensate for the windowing effect that tapers the numerator of eq. (1) towards zero for large lags. A nice by-product of the division is that it can mitigate impacts of strong formants [4, 6, 8].

Both cACF and sACF can be normalized by their value at k = 0, respectively. The normalized cACF has a range of [0, 1], while the normalized sACF does not guarantee this range. When the normalized sACF exceeds 1, it will be replaced by its reciprocal to satisfy the range requirement of [0, 1].

We use synthetic speech to assess the normalized cACF and sACF for pitch detection. The glottal source is simulated using the LF model [9], and the vocal tract response is approximated by the first formant. Three F0 values (63.5, 125, and 260 Hz) and two first formant frequencies (260 and 680 Hz) are considered to examine the impact of windowing and formant oscillation. These two formant frequencies roughly correspond to the lowest and highest first formant frequency in speech. As in [6], we find that sACF can sometimes over-compensate for the windowing effect, yielding PDA errors. Fig. 1 b) represents one such example where the overcompensation occurs at around 15 ms. Once taking the reciprocal, there appears a drop in the plot when k > 15 ms and hence, an erroneous pitch estimation.

A little thought convinces us to introduce an exponent to eq. (1) to achieve a good balance between cACF and sACF. We call the new variant the exponential ACF (eACF):

$$r_{xx,e}(t,k) = \frac{r_{xx,c}(t,k)}{[\sum_{j=0}^{N-1} w(j)w(j+k)]^{\beta}} \tag{2}$$

2.3. Integration with pitch continuity

In Section 2.1, the harmonicity, clarity, and periodicity are all estimated by looking at harmonics in individual frames. Consequently, SAD using these features tends to report more false alarms, i.e., mistaking a noise segment for a speech one in the presence of noise. To overcome this problem, pitch continuity in speech is explored and integrated here.

We assume that for voiced segments, the pitch remains approximately unchanged between two adjacent, short frames. Lets consider 3 consecutive frames. Assume that the index of the current frame is t and that the associated pitch resides at p_t (k_{max}, q_{min} or ω_{max}). Then we compute the corresponding values: $r_{xx,s}$, $AMDF$ and HPS for frames t-1, t, t+1 all using p_t. We call them corresponding values because they are directly corresponded to the pitch. For a speech segment, these values should be quite close to each other in continuous frames. On the other hand, the corresponding values of frame t-1 and t+1 at p_t may differ considerably from that of frame t for noise segments, owing to the absence of the continuity of the harmonics. This is illustrated in Fig. 2 on real-world data, where it can be seen that k_{max} vary dramatically among noise frames, but stay almost unchanged among speech frames.

Based on the above findings, we come up with the following method to integrate pitch continuity. We introduce the superscript ' to designate a feature that is based on pitch continuity and the subscript t in $k_{max}, q_{min} and \omega_{max}$ to designate the pitch of the frame t. New harmonicity, clarity, and periodicity are computed as follows:
1) *Harmonicity*:

$$H^{'}(t) = \underset{i=-1,0,1}{median}\{\frac{r_{xx,s}(t+i,\mathbf{k_{max,t}})}{r_{xx,s}(t+i,0) - r_{xx,s}(t+i,\mathbf{k_{max,t}})}\} \tag{3}$$

2) *Clarity*:

$$C^{'}(t) = \underset{i=-1,0,1}{median}\{\frac{1 - AMDF(t+i,\mathbf{q_{min,t}})}{AMDF(t+i,q_{max,t+i})}\} \tag{4}$$

3) *Periodicity*:

$$P^{'}(t) = HPS^{'}(t) = \underset{i=-1,0,1}{median}\{HPS(t+i,\mathbf{\omega_{max,t}})\} \tag{5}$$

It is important to note that the described method is quite different from a simple median filter that works directly on the final features. For instance, in Fig.2.(c), if we apply a normal median filter to it, the input values would be the red squares instead of the ones on the red dash. And because they are all at their maximums, a median filter won't decrease the value that much as we do now. Our method takes into account the position of those maximums, which is the pitch and thus has to be continuous among frames. We enable this by playing a trick of reversing the process. That is, we first work out the pitch from the current frame by finding the maximum/minimum, and then plug it into adjacent frames to obtain its corresponding values there. On the other hand, in the original method, only the maximum/minimum values are used for feature computation, ignoring the implied pitch, let alone its continuity.

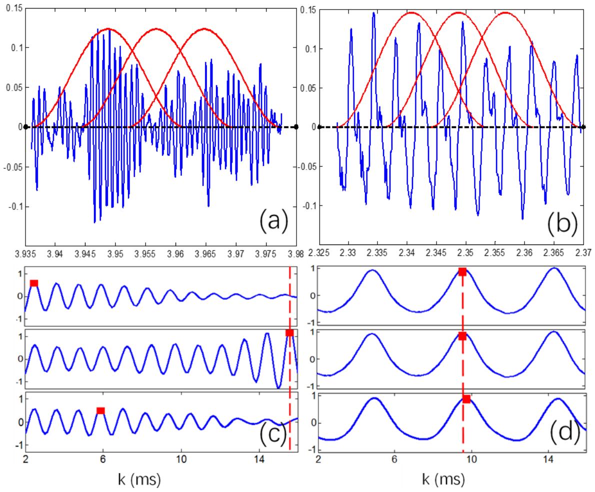


Fig. 2. Illustration of pitch continuity for (a) noise and (b) voiced speech. (c) and (d) show the normalized ACF of (a) and (b) respectively. The three curves in (c) and in (d) denote the left, the center, and the right frame (from top to bottom). The red squares denote the k_{max} of each frame. The red dash line denotes k_{max} of the middle frame for ease to compare.

It should also be noted that the harmonicity calculation is totally based on the harmonics-to-noise ratio of a frame, not the absolute energy of the harmonics. If the audio has drift noise or other slow-varying artifacts, the harmonicity calculation can be improved by first computing frame-wise zero crossing rates. If the zero crossing rate is below a preset threshold, we set the harmonicity for that frame to 0.

Finally, we adopt the approach in [4] for feature fusion and decision making, for a more straightforward comparison.

3. EXPERIMENTAL RESULTS AND DISCUSSION

The experiments are conducted on two databases. One is the training subset of the New England Region of the TIMIT database [10]. The other one is the Baihu database, a Chinese corpus we have collected for this study, in order to better examine the effectiveness of the pitch-continuity method on

type	0 dB		5 dB		10 dB		20 dB	
	A	B	A	B	A	B	A	B
office	9.08	17.05	6.24	12.98	4.65	7.30	3.19	3.47
volvo	12.67	19.22	7.28	12.95	4.68	7.91	3.58	3.63
factory	12.36	13.66	5.83	6.04	3.66	3.77	2.71	2.50
babble	16.56	12.38	9.17	6.87	5.54	4.23	3.07	2.60
white	2.84	2.45	2.68	2.27	2.79	2.18	2.80	2.19
subway	6.38	5.80	3.64	3.81	2.90	2.80	2.41	2.02
pooled	**10.23**	12.07	**6.15**	7.61	**4.22**	5.35	3.01	**2.98**

Table 1. Comparison of minDCF in % between the present method (Column A) and the Combo-SAD (Column B, our implementation) on the Baihu database.

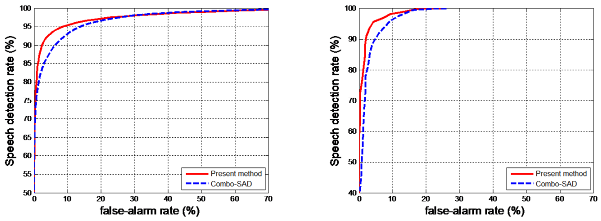
a tone language. The Baihu database is composed of speech from 10 male and 10 female speakers. Most of the speakers each speak 50 Chinese utterances, and each utterance contains only one speech segment. Two other female speakers contribute 80 and 100 utterances, respectively, and each utterance contains two well-separated speech segments. The actual starting and ending points of speech segments are manually labeled.

The noisy data is generated by adding noise data to the above clean speech, at a specified SNR by FaNT [11]. Six different types of noise are available: white, babble, factory, volvo (from the NOISEX-92 database [12]), subway noise (from the FaNT distribution), and finally an office background noise that we recorded [13]. The Baihu database is freely available upon request.

Error rates are estimated from the amount of time that is misclassified by the system, in the way as specified in the official NIST OpenSAD [1]. The only difference is that if the result shows a silence or pause between two talkspurts that is less than 0.4 seconds, we will smooth it as a voice part. The Detection Cost Function (DCF) in [1] will be used as the main metrics for performance evaluation.

Table 1 presents the minDCF results for 6 different noise types and at 4 different SNR levels on the Baihu database. The corresponding results of [4] (our own implementation) are also given. As shown in Table 1, the present method overall outperforms the original Combo-SAD method. Specifically, our method is able to significantly lower minDCF for volvo, factory, and office noise types. This is expected because pitch continuity is most effective in non-stationary or instantaneous types of noise. For the other 3 types, the Combo-SAD method is better. Recall that in the proposed method, we excluded prediction gain and perceptual spectral flux, while we maintained them in our implementation of the Combo-SAD. It is possible that by adding these two features, the result for the present method can be further improved.

In Table 1, we pool together all the test data from 6 noise types. It can be seen that the present method lowers minDCF across almost all SNR levels. For example, at SNR = 0 dB, the minDCF of the present method is 10.23%, a relative reduction of more than 15% when compared to that of the Combo-SAD.



(a) Baihu database

(b) TIMIT database

Fig. 3. Comparison of ROC curves between the present method and Combo-SAD on (a) aggregate data of 6 noise types and 4 SNR levels on Baihu database, and (b) the TIMIT database added by volvo noise at SNR = 5 dB.

We also aggregate all the data by 6 noise types and 4 SNR levels. Fig. 3 (a) depicts the Receiver Operating Characteristic (ROC) curves for this aggregated set of data. At the 3% false-alarm rate, as we can see, the miss rate is 8.5% for the present method and 14% for the Combo-SAD method, which pertains to an improvement of 39.3% relatively.

The result for the TIMIT database is provided in Fig. 3 (b), where only the ROC curve for the volvo noise and SNR = 5 dB is plotted, because of the interest in in-car applications.

4. CONCLUSION

In the above, we first reviewed Combo-SAD and the underlying ACF calculation of [6]. We used synthetic speech to illustrate the effects of windowing and tapering in ACF as lags increase. Compared to [6], we introduced an exponent parameter β in eq. (2). As shown in Fig. 1, use of β can effectively mitigate the windowing effect while preventing overcompensation from happening. We plan to optimize the value of β for accurate pitch detection.

We also propose a feasible way to integrating pitch continuity with harmonicis related features. Its effectness was evaluated by comparing the new features to the original ones on artificially generated noisy data. For the two databases (TIMIT and Baihu), the present method was advantageous in three noise types, and was comparable to the Combo-SAD method for the other three types. Overall, we achieved better performance, even though in our method we did not use the features of prediction gain and perceptual spectral flux.

We have started to collect real-world noisy speech data to evaluate our method and other methods. We will also explore how to gain access to OpenSAD data to augment our experiments (currently they are only available to the participants of the openSAD evaluation). Finally, the possibility of incorporating pitch continuity in deep neural nets is also worth investigation.

5. REFERENCES

[1] NIST, “Evaluation plan for the nist open evaluation of speech activity detection,” 2016.

[2] M. Sahidullah and G. Saha, “Comparison of speech activity detection techniques for speaker recognition,” *arXiv preprint arXiv:1210.0297*, 2012.

[3] M. Graciarena, A. Alwan, D. Ellis, H. Franco, L. Ferrer, J. H. Hansen, A. Janin, B. S. Lee, Y. Lei, V. Mitra, *et al.*, “All for one: feature combination for highly channel-degraded speech activity detection,” in *INTERSPEECH*, pp. 709–713, 2013.

[4] S. O. Sadjadi and J. H. Hansen, “Unsupervised speech activity detection using voicing measures and perceptual spectral flux,” *IEEE Signal Processing Letters*, vol. 20, no. 3, pp. 197–200, 2013.

[5] D. Joho, M. Bennewitz, and S. Behnke, “Pitch estimation using models of voiced speech on three levels,” in *ICASSP*, vol. 4, pp. IV–1077, IEEE, 2007.

[6] P. Boersma, “Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound,” in *Proceedings of the institute of phonetic sciences*, vol. 17, pp. 97–110, Amsterdam, 1993.

[7] L. R. Rabiner and R. W. Schafer, *Theory and applications of digital speech processing*, vol. 64. Pearson Upper Saddle River, NJ, 2011.

[8] E. Chuangsuwanich and J. Glass, “Robust voice activity detector for real world applications using harmonicity and modulation frequency,” in *Twelfth Annual Conference of the International Speech Communication Association*, 2011.

[9] G. Fant, J. Liljencrants, and Q.-g. Lin, “A four-parameter model of glottal flow,” *STL-QPSR*, vol. 4, no. 1985, pp. 1–13, 1985.

[10] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, N. L. Dahlgren, and V. Zue, “Timit acoustic-phonetic continuous speech corpus,” *Linguistic data consortium*, vol. 10, no. 5, p. 0, 1993.

[11] H. G. Hirsch, “Fant: filtering and noise adding tool,” *Niederrhein University of Applied Sciences, http://dnt.-kr. hsnr. de/download. html*, 2005.

[12] A. Varga and H. J. Steeneken, “Assessment for automatic speech recognition: Ii. noisex-92: A database and an experiment to study the effect of additive noise on speech recognition systems,” *Speech communication*, vol. 12, no. 3, pp. 247–251, 1993.

[13] Q. L. Y. Liu, J. Wang and S. Wang, “A novel speech activity detection algorithm based on the fusion of time domain and frequency domain features,” *Jiangsu University of Science Technology*, vol. 31, no. 01, pp. 73–78, 2017.