Significance of single frequency filter for the development of children’s KWS system

Biswaranjan Pattanayak and Gayadhar Pradhan

Department of Electronics and Communication Engineering, National Institute of Technology Patna, India

Abstract

Spotting a defined set of keywords from a running speech is known as keyword spotting (KWS). When keywords are detected using speech data from child speakers with the acoustic model built using speech data from adult speakers, it is named as childrens KWS system. Owing to the differences in pitch and speaking rate between the two kind of speakers, the performance of childrens KWS system deteriorates severely. In this paper, a pitch independent feature extraction method is proposed exploiting single frequency filtering (SFF) approach to address this issue. The method aims at finding the amplitude envelopes at Mel spaced frequencies. These amplitude envelopes are then averaged per analysis frame. Logarithm of the means are computed followed by Discrete Cosine Transform (DCT) to determine the required pitch robust feature, here denoted as Mel spaced single frequency filtering cepstral coefficient (MS-SFF-CC). The proposed feature outperforms several explored features with acoustic model trained on deep neural network-hidden Markov model (DNN-HMM) under pitch matched and mismatched test scenarios without and with data-augmented training.

Index Terms: KWS, single frequency filter, pitch, pitch robust feature

1. Introduction

Communication between humans and machines is termed as human-machine interaction. Machines react to the users through various user interfaces. Speech is one such user interface. Several speech-based human-machine interaction applications such as classifying calls and probing speech databases in call centres and security systems require a recognition vocabulary which may be small or large [1]. In such applications, automatic speech recognition (ASR) technology is employed to firstly convert the entire speech to text and then indexing and searching operations are executed. Spotting a defined set of keywords from a continuous speech is commonly known as keyword spotting (KWS). Over the years, several approaches are adopted for implementing KWS systems. Some of them include separating keywords and non-keywords employing self-organizing feature maps (SOFM) [2], using sub-word units such as phones, diphones, triphones to model keywords or filler models [3], feature space trace time normalization [4] and large vocabulary continuous speech recognition (LVCSR) [5, 6].

Mismatch in training and test scenarios is one of the major factors responsible for degrading the performance of the KWS system. In a KWS system, mismatch may happen due to out-of-vocabulary (OOV) words, channel distortion, use of foreign accent words, environmental noise, and speaker-dependent variations [7, 8, 9, 10]. Owing to the differences in pitch and speaking rate between the adult and child speakers, the performance of the KWS system deteriorates severely when the acoustic models are trained on adults’ speech and tested using children’s speech [11, 7, 12]. In literature, various methods are proposed to mitigate the mismatch in the development of children’s KWS systems. Some of those approaches include the reconstruction of smoothed spectra by applying variational mode decomposition (VMD) on short-term magnitude spectra [10, 7], adaptive low-pass filtering of short-term magnitude spectra [12], pitch adaptive cepstral truncation (PACT) [13] and pitch adaptive spectral normalization [14]. The acoustic feature like spectral moment time-frequency distribution augmented by low-order cepstral (SMAC) [15] is also explored for development of children’s KWS system. Recently, the data-augmented training through explicit prosody modification is explored for addressing pitch mismatch in children’s KWS system [11, 15, 12].

In the recent past, the single frequency filtering (SFF) approach is proposed in [16] to compute the amplitude envelope of the speech signal with very high spectral and temporal resolution. SFF features are used in applications such as separating speech and nonspeech regions (VAD) [16], replay attack detection [17] and extraction of epochs from telephone speech [18]. Pitch information in the speech signal is captured by the SFF features with high-frequency resolution. On the other hand, by determining the amplitude envelope at a particular frequency the pitch effect can be reduced. Motivated by this, we have presented a methodology to derive pitch robust acoustic features exploiting SFF. The KWS system is implemented using large vocabulary continuous speech recognition (LVCSR)-based approach [5, 6] employing deep neural network (DNN)-hidden Markov model (HMM) acoustic models [19]. The proposed feature is found to be superior to MFCC [20] and MFCC computed by spectral smoothing using VMD (VMD-MFCC) [10, 21] and PACT (PACT-MFCC) [13] with and without data-augmented training.

The rest of the paper is organized as follows: Section 2 illustrates the feature extraction procedure for the proposed approach. The details regarding the experimental setup for the development of the pitch robust KWS system are explained in Section 3. The experimental results are presented in Section 4. Finally, Section 5 concludes this study.

2. Proposed acoustic feature using SFF

As reported in [16], the SFF captures the instantaneous frequency components present in a speech signal. To achieve this, the input speech signal is first multiplied with a complex exponential which results in shifting the signal in frequency. Filtering of this frequency-shifted signal is then done with the help of SFF having the pole positioned very close to the unit circle at

© Copyright 2022 ISCA

Interspeech 2022
18-22 September 2022, Incheon, Korea

10.21437/Interspeech.2022-980

3183
2.1. SFF approach
The given speech signal is processed through the following sequence of steps to compute the instantaneous amplitude envelope using the SFF approach.

1. The first-order difference of the speech signal \( s[n] \) is firstly performed in order to discard any low-frequency components present in the signal,
   \[
   x[n] = s[n] - s[n - 1]
   \]  
   (1)

2. Then a complex exponential signal is multiplied with \( x[n] \) to shift the signal in frequency by \( f_p \) as follows:
   \[
   x[n, p] = x[n]e^{\frac{2\pi}{f_s}jn}
   \]  
   (2)
   where, \( f_p \) and the desired frequency \( f_p \) are related as,
   \[
   f_p = \frac{f_s}{2} - f_p
   \]  
   (3)

3. The frequency shifted signal \( x[n, p] \) is passed through the SFF whose transfer function is given by,
   \[
   H(z) = \frac{1}{1 + rz^{-1}}
   \]  
   (4)
   where, \( r \) is the location of the pole in the negative real axis of the z-plane whose value is set at 0.995 in order to make the filter stable. After filtering, the output signal \( y[n, p] \) is given by,
   \[
   y[n, p] = -ry[n - 1, p] + x[n, p]
   \]  
   (5)

4. Finally, the envelope of the SFF signal \( y[n, p] \) is calculated as,
   \[
   v[n, p] = \sqrt{(Re\{y[n, p]\})^2 + (Im\{y[n, p]\})^2}
   \]  
   (6)
   where, \( Re\{y[n, p]\} \) is the real part and \( Im\{y[n, p]\} \) is the imaginary part of \( y[n, p] \).

2.2. Extraction of acoustic feature employing SFF
Fig.1 illustrates the procedure for extracting the proposed acoustic feature employing SFF. Firstly, the speech signal is passed through a pre-emphasis filter to boost the high-frequency components. In the next step, the amplitude envelopes are determined at \( P \) frequency points separated in Mel scale following the procedure explained in Section 2.1, which results in producing \( P \) amplitude envelopes. These \( P \) frequency points are the same as the centre frequencies of the triangular Mel-filterbank employed for extracting standard MFCC feature [20]. Then the average of these envelopes is computed with a frame duration of 20 ms and frame repetition rate of 10 ms as follows,
   \[
   v_{avg}[k, p] = \frac{1}{2M + 1} \sum_{m=-M}^{M} v[n + m, p]
   \]  
   (7)
   where, \( p \) varies as 1, 2, 3, ..., \( P \) and \( k \) denotes a particular frame. The value of \( M \) is fixed at 160 for 16 kHz sampled speech data.

The average envelopes \( v_{avg}[k, p] \) are logarithmically compressed and DCT is applied to obtain a set of cepstral coefficients. The cepstral features computed by this process are termed as Mel spaced single frequency filtering cepstral coefficient (MS-SFF-CC). In this study, amplitude envelopes are computed at 40 Mel spaced points and the dimension of the base MS-SFF-CC feature is fixed at 13.

2.3. Effect of pitch variations on the MS-SFF-CC feature
To determine the effect of pitch variations on the proposed acoustic feature, the variance of the proposed MS-SFF-CC and conventional MFCC features are compared between two pitch group speakers. For this comparison, firstly TIMIT database [22] is divided into two groups, namely low-pitch (\( F_0 < 150 \) Hz) speaker groups and high-pitch (\( F_0 > 220 \) Hz) speaker groups. Then the variance is computed for two groups of speakers using the centre speech frames of the vowel /i/y/ using the valid reference markings available in the database. The variance is plotted for MFCC and MS-SFF-CC features in Fig 2. It can be seen that the variance is very low for the proposed MS-SFF-CC feature for both lower and higher feature indexes when compared with that of the MFCC feature. Therefore, the KWS system implemented employing the MS-SFF-CC feature will be less sensitive to pitch variations.

3. Experimental setup
3.1. Speech data
Two British English speech corpora namely the WSJCAM0 adults’ speech corpus [23] and the PF-STAR children’s speech corpus [24] are used for the experimental evaluations in this work. Both the speech corpora are noise-free, i.e. the recording environment is clean. Two entirely separated datasets are prepared from the WSJCAM0 adults’ speech corpus. One is designated as train_set_AD and the other one is test_set_AD. The train_set_AD is comprised of 13.3 hours of speech data with 6812 utterances from 80 speakers. Likewise, test_set_AD consists of 2.9 hours of speech data with 2990 utterances from 32 speakers. The train_set_AD is employed for building the
DNN-HMM-based acoustic models, whereas the test_set_AD is used for performance evaluation in matched test scenarios. Another dataset named test_set_CH is prepared from the PF-STAR children’s speech corpus. It consists of 6.5 hours of speech data with 642 utterances from 133 speakers. The test_set_CH is used for performance evaluation in mismatched test scenarios. A list of 10 keywords (zero, one, two, three, four, five, six, seven, eight, nine) and 20 keywords (zero, one, two, three, four, five, six, seven, eight, nine, people, there, bank, point, with, year, month, they, ten, number) which are repeated in both test_set_AD and test_set_CH are picked for evaluating the performance of KWS system under matched and mismatched test scenarios. The speech waves present in all the datasets are sampled at 16 kHz.

3.2. Feature extraction

Pre-emphasis of the speech signal is accomplished using a filter coefficient of 0.97. The short-time analysis is performed with overlapping Hamming windows having frame duration and frame shift of 20 ms and 10 ms, respectively. Extraction of the 13 dimensional base MS-SFF-CC feature is done following the procedure discussed in Section 2. The explored features such as MFCC [20], VMD-MFCC [10, 21] and PACT-MFCC [13] are derived using the method as explained in the original reported works. For all the explored feature extraction processes, the number of filters used in the filter-bank is 40 and the dimension of the base cepstral coefficient is set at 13. Time-splicing of all the base features is done taking into account a context size of 9 frames. The dimension of the feature vector is reduced to 40 by employing linear discriminant analysis (LDA) followed by the maximum likelihood linear transform. Finally, feature normalization is achieved with cepstral mean and variance normalization (CMVN) as well as feature-space maximum-likelihood linear regression (fMLLR) [25].

3.3. Building the acoustic models

In the present study, the KWS system is developed using Kaldi toolkit [26]. As stated previously, the KWS system is implemented using the LVCSR-based approach [11, 7]. There are two stages in the LVCSR-based approach [1] of detecting keywords. The first stage converts the input speech to text and generates the corresponding lattices. The second stage utilizes these lattices to determine the word-level indexes which are finally used to detect the required keywords. For building the acoustic model, tri-state left to right context-dependent HMM structure is adopted. The statistical model parameters are determined by implementing a decision tree-based state tying with the maximum number of tied states (Senones) kept at 2000. The necessary state transition probabilities for the HMM states are determined by GMM and DNN. 8 diagonal co-variance Gaussian components are used to model every HMM state in the GMM-HMM system. The training of the DNN-HMM system is initiated by using the alignments produced by the GMM-HMM system. The learning of the DNN-HMM system is achieved by time-splicing of the fMLLR-normalized feature vectors once again with a context size of 9 frames. The DNN-HMM system has 8 number of hidden layers and every layer is consisting of 1024 hidden nodes having tanh nonlinearity. In the beginning, the learning rate is set at 0.015 and finally, after reduction, it is fixed at 0.002.

4. Evaluation result and discussion

4.1. KWS metric

The metric used for evaluating the performance of the KWS system is the term-weighted value (TWV) [27]. Two variants of TWV’s are considered to evaluate the effectiveness of the proposed technique, namely actual term-weighted value (ATWV) and maximum term-weighted value (MTWV). ATWV represents the TWV at the actual decision threshold and MTWV denotes the TWV at the best threshold [28]. The two probabilities involved in the computation of TWV are probability of miss detection ($P_{miss}$) and probability of false alarm ($P_{fa}$). As presented in [28], for a particular keyword (kd) and spotting threshold (γ), firstly, the parameters are computed by,

$$ P_{miss}(kd, \gamma) = 1 - N_{ref}(kd, \gamma)/N_{eh}(kd) \quad (8) $$

$$ P_{fa}(kd, \gamma) = N_{fa}(kd, \gamma)/N_{ref}(kw) \quad (9) $$

$$ TWV(kd, \gamma) = 1 - P_{miss}(kd, \gamma) - \beta P_{fa}(kd, \gamma) \quad (10) $$

where, $N_{ref}(kd, \gamma)$ is the number of instances when the keyword kd is correctly detected using a system dependent threshold $\gamma$, $N_{eh}(kd)$ represents the number of reference happenings of the keyword kd, $N_{fa}(kd, \gamma)$ specifies the number of times the keyword kd is falsely spotted at threshold $\gamma$. $N_{ref}(kw)$ is the number of cases when the keyword kd is off-target and $\beta$ is a constant fixed at 999.9.

Finally, $P_{miss}$, $P_{fa}$ and TWV at a threshold $\gamma$ are calculated by taking the mean of $P_{miss}(kd, \gamma)$, $P_{fa}(kd, \gamma)$ and $TWV(kd, \gamma)$, respectively, as follows:

$$ P_{miss}(\gamma) = \frac{1}{T} \sum_{kd=1}^{T} P_{miss}(kd, \gamma)/T \quad (11) $$

$$ P_{fa}(\gamma) = \frac{1}{T} \sum_{kd=1}^{T} P_{fa}(kd, \gamma)/T \quad (12) $$

$$ TWV(\gamma) = \frac{1}{T} \sum_{kd=1}^{T} TWV(kd, \gamma)/T \quad (13) $$

where $T$ is the total set of keywords. A better KWS system requires TWV to be higher (close to 1) [27]. Conversely, lower values of $P_{miss}$ and $P_{fa}$ are preferred.

4.2. Performance of MS-SFF-CC feature in pitch matched and mismatched test scenarios

As mentioned earlier, the training of the acoustic models is achieved using train_set_AD. For testing the KWS system un-
under the matched scenario, test_set_AD is used. Likewise, test_set_CH is used for testing under the mismatched scenario. The performances of the KWS systems in terms of MTWV, ATWV, $P_{miss}$, and $P_{fa}$ are summarized in Table 1 for the defined list of 10 and 20 keywords. It has been observed that both types of TWV’s are very low for the test_set_CH as compared with the test_set_AD. This is due to the difference in pitch and speaking rate between the adult and child speakers. However, improved performance is observed for the proposed feature as compared to the explored features under matched and mismatched test scenarios for 10 keywords as well as 20 keywords. But the improvement is significant for the test_set_CH. Hence, the effect of pitch variations is less on the proposed feature MS-SFF-CC when compared with the explored features for the identical task.

Table 1: Performance of the developed KWS system under pitch matched (test_set_AD) and mismatched (test_set_CH) test scenarios. The performance is given in terms of $P_{miss}$, $P_{fa}$, ATWV and MTWV employing proposed and explored features.

<table>
<thead>
<tr>
<th>Fac.</th>
<th>Pmiss</th>
<th>Pfa</th>
<th>ATWV</th>
<th>MTWV</th>
<th>Pmiss</th>
<th>Pfa</th>
<th>ATWV</th>
<th>MTWV</th>
</tr>
</thead>
<tbody>
<tr>
<td>MFCC</td>
<td>0.005</td>
<td>0.004</td>
<td>0.061</td>
<td>0.108</td>
<td>0.006</td>
<td>0.005</td>
<td>0.108</td>
<td>0.108</td>
</tr>
<tr>
<td>VMD-MFCC</td>
<td>0.053</td>
<td>0.030</td>
<td>0.064</td>
<td>0.200</td>
<td>0.053</td>
<td>0.030</td>
<td>0.064</td>
<td>0.200</td>
</tr>
<tr>
<td>PACT-MFCC</td>
<td>0.010</td>
<td>0.010</td>
<td>0.061</td>
<td>0.108</td>
<td>0.010</td>
<td>0.010</td>
<td>0.061</td>
<td>0.108</td>
</tr>
<tr>
<td>MS-SFF-CC</td>
<td>0.010</td>
<td>0.010</td>
<td>0.061</td>
<td>0.108</td>
<td>0.010</td>
<td>0.010</td>
<td>0.061</td>
<td>0.108</td>
</tr>
</tbody>
</table>

4.3. Impact of data-augmented training on the proposed and explored features

Earlier reported works [11, 7] suggest that the data-augmented training through explicit prosody modification is an effective approach for improving the performance of children’s KWS system. In this study, the pitch modified train_set_AD dataset is created following the approach presented in [11, 7, 29]. The optimal values of pitch scaling factor (PS) are determined and features are extracted from the PS-modified speech data. These features along with the features extracted from the original speech data (train_set_AD) are pooled together to build the DNN-HMM-based acoustic models. The performance is evaluated employing the original test data. The performances of the proposed and explored features with respect to data-augmented training for two optimal values of PS (PS125 and PS135) are given in Table 2. The performances are given in terms of MTWV, ATWV, $P_{miss}$ and $P_{fa}$ for a defined set of 10 and 20 keywords. It can be observed that the best values of TWV’s are obtained for the proposed feature as compared to all the explored features for 10 keywords as well as 20 keywords. Hence, the proposed feature MS-SFF-CC is superior compared to the explored features.

5. Conclusion

In this paper, to address the acoustic mismatch due to pitch variations a pitch robust feature extraction technique employing SFF is proposed, which is termed as MS-SFF-CC. An effective children’s KWS system is implemented utilizing the MS-SFF-CC feature. The MS-SFF-CC feature provides better performance in terms of both MTWV and ATWV when compared with the features like MFCC, VMD-MFCC and PACT-MFCC reported for an identical task under matched and mismatched test scenarios. Moreover, training with data-augmentation is studied using MS-SFF-CC and explored features. With the use of data-augmented training, the performance of the developed KWS system is further improved.

6. Acknowledgments

The authors wish to declare that we have not received any financial support from any source for the accomplishment of this work.
7. References


