Improvement of Audio-Visual Keyword Spotting System Accuracy Using Excitation Source Feature

Salam Nandakishor(✉) and Debadatta Pati

Department of Electronics and Communication Engineering, National Institute of Technology Nagaland, Chumukedima, Dimapur 797103, Nagaland, India
salamnandu@gmail.com

Abstract. In this paper, we proposed a robust audio-visual keyword spotting (AVKS) system. This system is developed using DNN (Deep Neural Network) model with State Level Minimum Bayes Risk (sMBR) criteria. The symbols of International Phonetic Alphabet (IPA) are used for representing the speech sounds at phonetic level. Our proposed system can recognize 34 phonemes, silence region and can also detect the predefined keywords formed by these phonemes. Most of the audio-visual keyword spotting system used Mel-frequency cepstral coefficient (MFCC) as audio feature. This feature represents only the vocal-tract related information but does not contain excitation source information. Therefore, we explore the excitation source features as the supplementary information in this work. The excitation source features extracted from glottal flow derivative (GFD) and linear prediction (LP) residual through standard mel cepstral analysis are termed as Glottal Mel-Frequency Cepstral Coefficient (GMFCC) and Residual Mel-Frequency Cepstral Coefficient (RMFCC) respectively. The GFD signal is generated using Iterative Adaptive Inverse Filtering (IAIF) method whereas LP residual is estimated by inverse filtering process. In our experimental analysis, we observe that the performance of glottal based excitation feature is better than LP residual based excitation source feature in keyword spotting task. Hence, we consider the GMFCC features in development of our proposed system. The AVKS system using MFCC and DCT (Discrete Cosine Transform) based visual features extracted from mouth region provides an average accuracy of 93.87%, whereas the inclusion of GMFCC feature improves the performance to 94.93%. The experimental observations show the benefit of excitation source information for audio-visual keyword spotter under noisy condition.

Keywords: AVKS · Residual mel-frequency cepstral coefficient · Glottal mel-frequency cepstral coefficient · State level minimum bayes risk

1 Introduction

Keyword spotting system enables to detect the specific words from the continuous speech. This type of system can be used in many real time applications

A. Karpov et al. (Eds.): SPECOM 2023, LNAI 14339, pp. 344–356, 2023.
https://doi.org/10.1007/978-3-031-48312-7_28

such as voice assistant, customer care service and smart speakers etc. However, such applications may not be working properly in noisy environment due to quality degradation of speech sound. This creates the difficulty in detection of keywords by the machine. To solve this problem, researchers use the visual features as complementary information along with audio feature for spotting the predefined keywords. This type of system is known as audio-visual keyword spotting (AVKS) system [1]. In this work, we developed a phoneme based audio-visual keyword spotting system. The MFCC audio feature is concatenated with DCT based visual feature by using feature level fusion approach which is known as early integration fusion [2]. As the rapid growth of machine learning techniques and availability of high computational machines, researchers used various neural networks approaches to improve the performance of keywords spotting systems [3–5]. Therefore, we employed DNN (Deep Neural Network) classifier for modeling the phonemes. We further improves the performance of the proposed system by using State Level Minimum Bayes Risk (sMBR) criteria. The main advantage of phoneme based audio-visual keyword spotter is that it can detect any keywords formed by combination of phonemes.

Speech production can considered as a process in which the time-varying vocal-tract system is excited by a time-varying excitation source [6]. Researchers used the audio feature MFCC for representing the vocal-tract characteristics in speech recognition tasks [7]. Some of speech sound units (phonemes) such as (/b/ and /p/, /t/ and /d/) have same place of articulation and same manner of articulation [8]. It means they are having similar vocal-tract characteristics. Moreover, these phonemes have similar lips shape or lips movements while producing the speech sounds [9]. Because of these similar vocal-tract characteristics and similar lips movements, it may create confusion in phoneme recognition and resulting inaccurate keyword detection from continuous speech. In this case, the supplementary feature; the excitation source information may be helpful to differentiate between similar sound units or phonemes by machine. The excitation source signal can be represented by linear prediction (LP) residual [10,11] or the glottal flow derivative (GFD) signal derived from speech signal. Several GFD extraction methods have been reported in [12–18]. In a recent work [19], the Iterative Adaptive Inverse Filtering (IAIF) approach is found effective for deriving the excitation source information from speech signal. In this work, GFD signal is computed by using IAIF algorithm and LP residual of speech signal is estimated by using inverse filtering approach.

The rest of the paper is organized as follows: Sect. 2 mentions some research works related to audio-visual keyword spotting task. Section 3 provides the details information about the database used in system development. Section 4 describes about the development of DNN-sMBR based audio-visual Keyword Spotting system. Experimental results analysis are discussed in Sect. 5. The conclusion of the paper is declared in Sect. 6.

2 Related Work

Very few research has been done to detect the keywords audio-visually [1]. In [20], authors developed an HMM based audio-visual keyword spotting system. The normalized histogram intensity of mouth region was used as visual feature. The MFCC audio feature was combined with visual feature using feature level fusion approach. The objective of proposed keyword spotter was to detect the 19 English keywords. The performance of this system was analyzed at various SNR levels. The noisy audio speech signals were generated by adding white noise. The audio-visual keyword spotter performed better than the audio based keyword spotter at all SNR levels.

In another work [21], authors used the conventional HMM garbage model in development of Mandarin based audio-visual keyword spotting system. They proposed a visual feature named as discriminative local spatial-temporal descriptor (disCLBP-TOP). The models built by acoustic feature MFCC and visual features were combined by adaptive integration approach with appropriate weights. A sigmoid function was used to generate these weights. The 30 Mandarin keywords belong to 12 male and 8 female speakers were used for system performance analysis. They also compared the performance between bimodal (audio-visual keyword spotter) and unimodal (audio or visual based keyword spotter) using white and babble noise added noisy speech signals.

In [1], authors presented a novel lip descriptor that comprise of both geometric features and appearance based features. The geometric features extracted from lips region were combined with appearance based spatiotemporal features. The audio features were extracted using mel cepstral analysis. Authors used two-step strategy HMM based keyword spotting system to make system more robust. At first stage, the acoustic and visual keyword with log-likelihoods were generated. The decision fusion was applied in the second stage to generate the final keyword. The OuluVS and PKU-AV database were used for experimental results analysis.

Table 1. The reported AVKS systems and their results.

Method	Accuracy (in %)
HMM based AVKS [20]	78
HMM-garbage based AVKS [21]	75.1
Two step strategy HMM based AVKS [1]	80.5

The results of the reported AVKS system are shown in Table 1. All of these reported works used the MFCC as audio feature in development of AVKS system. They were mainly focus on visual features to improve the system performance. It is interesting to explore the excitation source based audio feature as supplementary information for audio-visual keyword spotting task.

3 Database Description

Our proposed system is developed using train data set of track 1 of the 2nd 'CHiME' Challenge audio database [25]. The utterance structure of this database is shown below.

[command(4)] [color(4)] [preposition(4)] [letter(25)] [digit(10)] [adverb(4)]

The numerical value shown inside the bracket defines the number of different commands, colors, prepositions, letters, digits and adverbs present in the database. The commands are BIN, LAY, PLACE and SET. The different colors available in the utterances are BLUE, GREEN, RED and WHITE. Prepositions present in database are AT, BY, IN and WITH. The English alphabets from A-Z excluding W and digits from 0 to 9 are also available in the database. The words; AGAIN, NOW, PLEASE and SOON are the 4 adverbs utterances used in this database. The database belongs to 34 speakers (18 male, 16 female) [25].

Clean audio speech material was taken from the Grid corpus [26]. The clean audio speech were convolved with a set of binaural room impulse responses ((BRIRs) to simulate the speaker movements and reverberation. The background noise recorded from the living room were mixed with the audio speech signals to generate the noisy speech signals. These noisy speech signals were generated at six different SNRs (-6 dB, -3 dB, 0 dB, 3dB, 6 dB and 9 dB).

The track 1 of the 2nd 'CHiME' database comprises of 3 data sets. They are (1) training set (2) development set and (3) test set. Each speaker of training set has 500 utterances. The development data set consists of 600 speech utterances at each SNR level. Similarly test data set contain same number of utterances of development data set. The speech signals were recorded with 16 bits and sampled at 16 kHz. The video features are extracted from video data of Grid database.

4 Development of DNN-sMBR Based Audio-Visual Keyword Spotting System

The processing steps involved in development of our proposed DNN-sMBR based audio-visual keyword spotting system are (a) Data Preparation and Pronunciation dictionary (b) Feature Extraction (c) Modeling and (d) Keyword Decoder.

Data preparation is an important step required in development of this proposed system. This processing step provides information about; (a) mapping of each speaker ID to its corresponding utterances IDs (b) paths of audio speech wav files along with their corresponding utterance IDs (c) assignment of each utterance ID with corresponding speaker ID and (d) transcription of all the utterances. Pronunciation dictionary contains the information about phonemes present in the database used for system development, silence region information and lexicon. The silence region is denoted by word 'SIL'. In order to understand the speech sounds in phonetic level, we analyze the mentioned 2nd 'CHiME' database properly and represent the phonemes by using International Phonetic

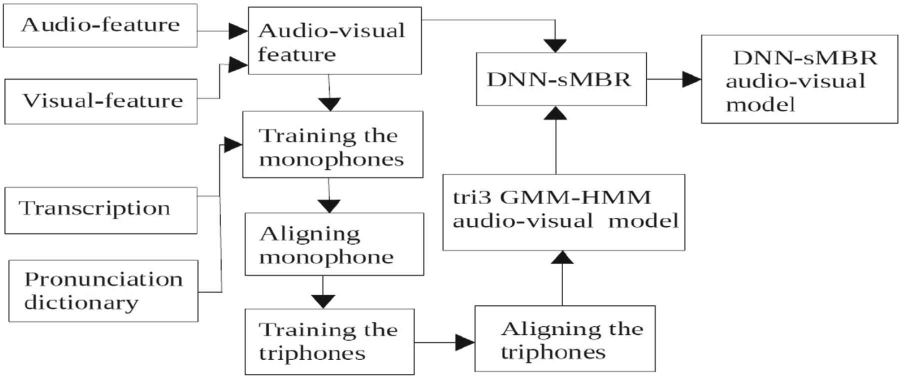

Fig. 1. Training phase of proposed audio-visual keyword spotting system.

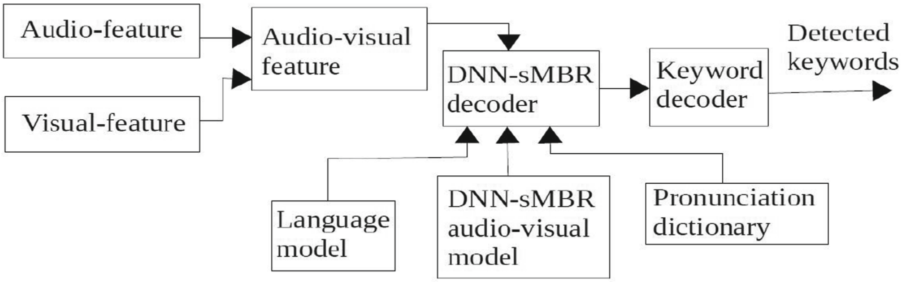

Fig. 2. Testing phase of proposed audio-visual keyword spotting system.

Alphabet (IPA) symbols. The IPA symbols and corresponding assigned ASCII of these phonemes are listed in Table 2.

The MFCC features are extracted from the input audio speech using standard mel cepstral analysis approach. The glottal excitation source based feature; GMFCC feature are derived from the GFD signal. These two features are concatenated to obtain a feature combination of vocal-tract and excitation source information. The dimension of the GMFCC is considered same as the 13 dimensional MFCC feature. Therefore, the total dimension of combined audio feature is 26. Image frames are extracted from the input video, then landmark points of mouth region are detected using Viola-Jones algorithm [27]. Discrete cosine transform (DCT) coefficients of landmark points of gray scale mouth region images are calculated to generate 63-dimensional visual features. In order to maintain same number of frames for both audio and visual features, the DCT coefficients are interpolated by using differential digital analyzer. Then the audio and visual features are concatenated in frame-wise manner to acquire 89 dimensional audio-visual features.

The block diagrams of training and testing phase of proposed audio-visual keyword spotting system are shown in Figs. 1 and 2. An audio-visual based

Table 2. List of phoneme used in development of audio-visual keyword spotting system.

Sl. No	Phoneme	Name in ASCII	Sl. No	Phoneme	Name in ASCII
1	b	B	18	w	W
2	ʃ	CH	19	y	Y
3	d	D	20	z	Z
4	ð	DH	21	aː	AA
5	f	F	22	æ	AE
6	g	G	23	ʌ	AH
7	ʤ	JH	24	ɔː	AO
8	k	K	25	ə	AX
9	l	L	26	ε	EX
10	m	M	27	i	IH
11	n	N	28	iː	IY
12	p	P	29	uː	UW
13	r	R	30	aʊ	AW
14	s	S	31	aɪ	AY
15	t	T	32	eɪ	EY
16	θ	TH	33	əʊ	OW
17	v	V	34	ɪə	IA

monophone GMM-HMM model is built using the audio-visual features, transcriptions and pronunciation dictionary. Contextual information of neighbouring phonemes that is front and back phoneme are not considered in monophone model. The audio-visual features are aligned with corresponding reference transcriptions using force alignment Viterbi algorithm. The monophone model is further extended to triphone GMM-HMM model. In this context-dependent triphone model, audio-visual features of neighbouring frames that is ± 3 frames are spliced to capture the dynamic information. Then, the dimension of this spliced audio-visual feature is reduced to 40 using Linear Discriminant Analysis (LDA) [28]. A popular speaker normalization technique, Maximum Likelihood Linear Transform is used to minimize the speakers variation. To make the proposed system more robust and speaker independent; speaker adaptation training (SAT) and feature-space Maximum Likelihood Linear Regression (fMLLR) [29] are used. This type of model is generally known as tri3 GMM-HMM audio-visual model. The pre-training of DNN is done by training the stack of Restricted Boltzmann machine (RBMs) through Contrastive Divergence (CD) approach. Updating of weights during the pre-training stage are used to initialize the DNN parameters, it allows the discriminative fine-tuning and reduce over fitting. During fine-tuning of DNN, the parameters are updated in a layer-wise manner by using back-propagation and Stochastic Gradient Descent (SGD) techniques. The sequence discriminative training method "state-level minimum Bayes risk (sMBR)” [30] is employed to emphasize the state sequence with better frame accuracy with respect to reference alignment. This model is represented as DNN-sMBR model. The model consists of 6 hidden layers with sigmoid activation function and used 18 beam for decoding. Each layer of DNN has 2048 neurons. At audio-visual keyword decoder stage, the system is ready to generate automatic transcriptions and detect the keywords of unknown test utterances. The system needs test input audio-visual feature, trained audio-visual model, language model and pronunciation dictionary to generate automatic transcription.

5 Experimental Results and Discussion

Visual keyword spotting systems are developed using different modeling approaches and compare their accuracies to select the best model for proposed AVKS system. We also compare the performance of LP based excitation source feature and glottal based excitation feature in context of keyword spotting task. Then, the outperforming model and excitation source feature are used in development of audio-visual keyword spotting system.

5.1 Visual Keywords Spotting System

Visual keyword spotting (VKS) is an automatic process of identifying the query keywords present in the video sequences using visual features. The 63-dimensional DCT coefficients visual features are extracted from mouth region images. These image frames are obtained from videos data of Grid database. The performance of the proposed visual keyword spotting system is evaluated using different models; GMM-HMM, DNN, DNN-sMBR-1 and DNN-sMBR-5 with visual features. The DNN-sMBR-1 and DNN-sMBR-5 represent the DNN model with state-level minimum Bayes risk (sMBR) criteria with 1 and 5 iterations respectively.

Table 3. Performance comparison of models for visual digits keyword spotting system.

Model	Accuracy (in %)
GMM-HMM	82.65
DNN	87.07
DNN-sMBR-1	88.27
DNN-sMBR-5	**89.29**

Digits are the most commonly used one time password (OTP) and input command words in real time applications like customer care service or automatic login system. Therefore, we consider the digits from zero to nine as the keywords for both visual keyword spotting system as well as audio-visual keyword spotting system. The development data set of Grid database is used for performance analysis of this visual keyword spotting system. From the experimental results provided in Table 3, we notice that DNN based system gives better accuracy than GMM-HMM based VKS system. The performance of DNN based system increases when the sMBR criteria is applied. We also analyze the performance of DNN-sMBR based system with different iterations, we observe the system performance improves when the number of iterations of sMBR training increase from 1 to 5. After 5^{th} iteration, no further improvement in the system performance. Therefore, we adopt DNN-sMBR model with 5 iterations in development of proposed AVKS system.

5.2 Selection of Excitation Source Feature for Proposed AVKS System

The excitation source features have been explored for phoneme recognition task. The use of excitation source feature in speech recognition area is very limited as compare to other application like speaker recognition task. Some of the works related to excitation source features are reported in [7,22–24]. They extracted the excitation source features from LP residual signal derived from speech signal. These excitation source features were used for improving the performance of phone recognizer. In this work, we explore the glottal based excitation source feature (GMFCC) and compared to the LP residual based excitation source feature (RMFCC).

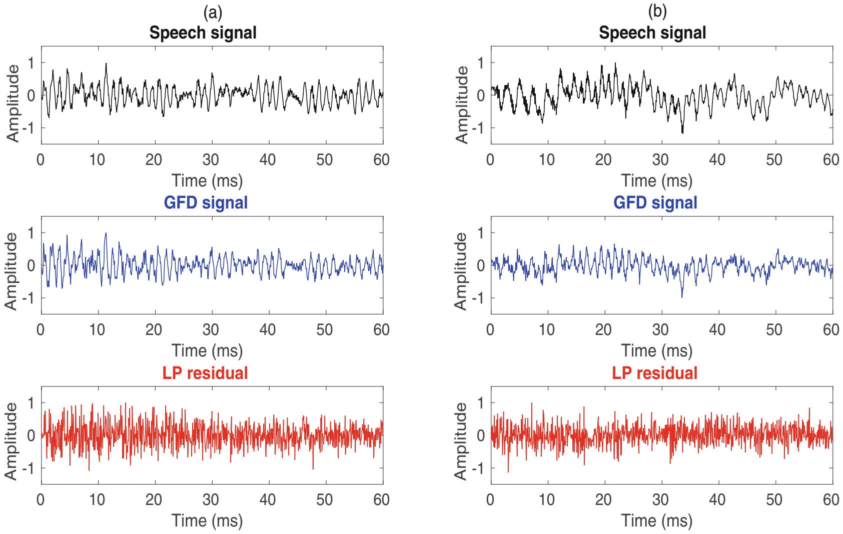

Fig. 3. Speech signal, GFD signal and LP residual of Alveolar Plosive sound units (a) t and (b) d.

Some of sound units such as (/p/ and /b/), (/t/ and /d/), (/k/ and /g/) are very confused among each other due to similar characteristics. We manually extract the phoneme /t/ portion from speech sound of English letter ‘T’ occur in utterance (srwt1n.wav) using linguistic tool ‘Praat’. Similarly the phoneme /d/ is extracted from sound ‘D’ present in utterance (lrwd1n.wav). The speech signal, GFD signal and LP residual of these Alveolar plosive sound units are shown in Fig. 3. The closeness between these Alveolar plosive sound units is analyzed by plotting kernel density using GMFCC and RMFCC feature. Kernel density estimation (KDE) method estimates the probability density function of feature vectors using kernels as weights and smoothing the density function by appropriate bandwidth. In this work, we used Gaussian kernel and bandwidth equal to 1.8 for plotting the kernel density of Alveolar plosive sound units.

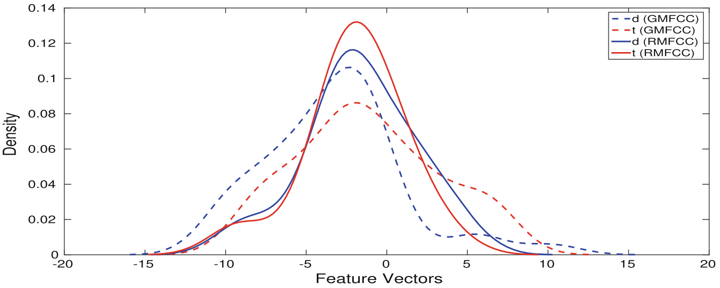

Fig. 4. The kernel density plot of Alveolar plosive sound units t and d using GMFCC and RMFCC feature.

In Fig. 4, blue color lines represent kernel density plots for sound unit /d/ and red color lines for phoneme /t/. The dotted lines belong to kernel density plots of GMFCC features whereas the solid lines is for RMFCC features. The dotted lines are less overlapping as compared to solid lines. This shows that GMFCC feature has more discriminative ability than RMFCC feature to distinguish the similar sound units like alveolar plosive. However, this type of analysis is a statistical approach. We further evaluate the performance of GMFCC and RMFCC feature in terms of keyword detection rate.

Table 4. Performance comparison of GMFCC and RMFCC at different SNR levels.

Feature	**−6 dB**	−3 dB	0 dB	3 dB	Average
GMFCC	68.73	76.12	88.66	92.27	**81.45**
RMFCC	61.34	63.75	77.84	83.16	71.45

The average accuracy of DNN-sMBR-5 based keyword spotter using GMFCC feature is 81.45% whereas RMFCC feature gives 71.45%. At all SNR levels, the glottal based excitation source feature performs better than LP residual based excitation source feature for detecting keywords. The experimental results available in Table 4 reveal that the GMFCC feature are robust than RMFCC under noisy condition. Therefore, we consider this glottal based excitation source feature in development of proposed AVKS system.

5.3 Performance Analysis of Proposed DNN-sMBR-5 Based Keyword Spotting System

The accuracies of digits keyword spotting using vocal-tract feature (MFCC), glottal based excitation source feature (GMFCC), visual feature (V) and their combination are given in Table 5. At 3dB, the accuracy of keyword detection using MFCC feature is more than GMFCC feature as well as visual feature. However, the performance degraded from 0 dB to -6 dB due to background noise present in utterances of test data set. Similar problem also affects the performance of GMFCC feature as well.

Table 5. Accuracies of Digits keyword spotting system with different features and its feature combination.

Feature	−6 dB	−3 dB	0 dB	3 dB	Average
MFCC	76.80	84.71	92.96	95.25	87.43
GMFCC	68.73	76.12	88.66	92.27	81.45
V	89.29	89.29	89.29	89.29	89.29
MFCC + GMFCC	79.21	86.94	93.30	96.05	88.88
MFCC + V	92.00	93.81	94.19	95.50	93.87
GMFCC + V	91.07	92.44	92.44	93.64	92.40
MFCC + GMFCC + V	93.99	94.50	94.85	96.39	**94.93**

From experimental results analysis, we know that MFCC feature is better than GMFCC for keyword spotting task. However, the combination of these features perform better than individual feature; MFCC and GMFCC. At all SNR levels, the accuracies of keyword spotting using MFCC feature are improved by combining with GMFCC feature. This shows the glottal excitation source features can be used as supplementary feature along with MFCC feature for automatic transcription and keyword spotting under noisy environment. The performance improvement in keyword spotting when added excitation source information to vocal-tract information is because of confusion reduction between similar phonemes.

The audio features can be corrupted by acoustic background noise, therefore visual feature are combined together with them. The performance of audio-visual keyword spotting system is not that much fluctuate as compare to audio-based keyword spotting system. The audio-visual keyword spotting system performs well at all SNR levels. We compare the performance of audio-visual keyword spotting systems developed by using (MFCC with visual feature), (GMFCC with visual feature), and combination of MFCC with GMFCC and Visual feature. The average accuracy of keywords spotting of these three audio-visual systems are 93.87%, 92.40% and 94.93% respectively. The AVKS system develop using MFCC along with visual feature gives better accuracy than system develop using GMFCC and visual feature together. The best average accuracy of keyword spotting is observed when MFCC, GMFCC and visual feature are combined together. These experimental results reveal that glottal excitation source feature is useful to use as supplementary information for audio-visual keyword spotting system under noisy environment.

6 Conclusion

In this work, we explore a glottal based excitation source feature (GMFCC), particularly for detecting the audio-visual keywords under noisy background. This glottal based excitation source feature is found more suitable than RMFCC feature for keyword spotting task. The performance of DNN based visual keyword spotting system is improved by using 'sMBR' sequence discriminative training with 5 iterations. This DNN model also works very well for audio-visual keyword spotting system. The feature combination of MFCC and GMFCC gives better accuracy than individual features; MFCC and GMFCC in keywords detection task. The performance of these combined audio feature degraded under noisy condition. To solve this issue, a robust visual feature is combined to audio features. The AVKS system developed using MFCC and visual features is improved by adding the glottal excitation source feature. In future work, the performance of proposed system can be evaluated by using unseen speakers data.

References

1. Pingping, W., et al.: A novel lip descriptor for audio-visual keyword spotting based on adaptive decision fusion. IEEE Trans. Multimedia **18**(3), 326–338 (2016)
2. Nandakishor, S., Pati, D.: Analysis of lombard effect by using hybrid visual features for ASR. In: Pattern Recognition and Machine Intelligence (PReMI 2021) (2021)
3. Higuchi, T., Gupta, A., Dhir, C.: Multi-task learning with cross attention for keyword spotting. In: IEEE Automatic Speech Recognition and Understanding Workshop (2021)
4. Berg, A., Connor, M., Cruz, M.T.: Keyword transformer: a self-attention model for keyword spotting. In: Proceedings of INTERSPEECH (2021)
5. Li, Y., et al.: Audio-visual keyword transformer for unconstrained sentence-level keyword spotting. In: CAAI Transactions on Intelligence Technology (2023)
6. Rabiner, L.R., Juang, B.-H., Yegnanarayana, B.: Fundamentals of Speech Recognition. Pearson Education (2012)
7. Manjunath, K., Rao, K.S.: Source and system features for phone recognition. Int. J. Speech Technol. **18**(2), 257–270 (2015)
8. International Phonetic Association: Hand Book of the International Phonetic Association, Cambridge University Press (1999)

9. Bear, H.L., Harvey, R.: Phoneme-to-viseme mappings: the good, the bad, and the ugly. Speech Commun. **95**, 40–67 (2017)
10. Yengnanarayana, B., Murthy, P.S.: Enhancement of reverberant speech using LP residual signal. IEEE Trans. Speech Audio Process. **8**(3), 267–281 (2000)
11. Prasanna, S.R.M., Srinivasa, C., Yengnanarayana, B.: Extraction of speaker-specific excitation information from linear prediction residual of speech. Speech Commun. **48**(10), 1243–1261 (2006)
12. Drugman, T., et al.: Detection of glottal closure instants from speech signals: a quantitative review. IEEE Trans. Audio Speech Lang. Process. **20**(3), 994–1006 (2012)
13. Naylor, P.A., et al.: Estimation of glottal closure instants in voiced speech using the DYPSA algorithm. IEEE Trans. Audio Speech Lang. Process. **15**(1), 34–43 (2007)
14. Thomas, M.R., Gudnason, J., Naylor, P.A.: Estimation of glottal closing and opening instants in voiced speech using the YAGA algorithm. IEEE Trans. Audio Speech Lang. Process. **20**(1), 82–91 (2012)
15. Murthy, K.S.R., Yegnanarayana, B.: Epoch extraction from speech signals. IEEE Trans. Audio Speech Lang. Process. **16**(8), 1602–1613 (2008)
16. Prathosh, A., Ananthapadmanabha, T., Ramakrishnan, A.: Epoch extraction based on integrated linear prediction residual using plosion index. IEEE Trans. Audio Speech Lang. Process. **21**(12), 2471–2480 (2013)
17. Alku, P.: Glottal wave analysis with pitch synchronous iterative adaptive inverse filtering. Speech Commun. **11**(23), 109–118 (1992)
18. Alku, P., Vilkman, E.: A comparison of glottal voice source quantification parameters in breathy, normal and pressed phonation of female and male speakers. IEEE Trans. Audio Speech Lang. Process. **48**(5), 240–254 (1996)
19. Dutta, K., Singh, M., Pati, D.: Detection of replay signals using excitation source and shifted CQCC features. Int. J. Speech Technol. **24**(9), 497–507 (2009)
20. Liu, M., et al.: Audio visual word spotting. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), vol. 3, pp. 785–788 (2004)
21. Liu, H., et al.: Audio-visual keyword spotting for mandarin based on discriminative local spatial-temporal descriptors. In: International Conference on Pattern Recognition, pp. 785–790 (2014)
22. He, J., Liu, L., Palm, G.: On the use of residual cepstrum in speech recognition. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), vol. 1, pp. 5–8 (1996)
23. Chengalvarayan, R.: On the use of normalized LPC error towards better large vocabulary speech recognition systems. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (1998)
24. Tripathi, K., Rao, K.S.: Improvement of phone recognition accuracy using speech mode classiffication. Int. J. Speech Technol. **21**(3), 489–500 (2018)
25. Vincent, E., et al.: The second "CHiME" speech separation and recognition challenge: datasets, tasks and baselines. In: International Conference on Acoustics, Speech, and Signal Processing (ICASSP), pp. 126–130 (2013)
26. Cooke, M., et al.: An audio-visual corpus for speech perception and automatic speech recognition. J. Acoust. Soc. Am. **120**(5), 2421–2424 (2006)
27. Ephraim, T., Himmelman, T., Siddiqi, K.: Real-time viola-jones face detection in a web browser. In: Canadian Conference on Computer and Robot Vision, pp. 321–328 (2009)
28. Rath, S.P., et al.: Improved feature processing for deep neural networks. In: Proceedings of the INTERSPEECH, pp. 109–113 (2013)

29. Povey, D., Saon, G.: Feature and model space speaker adaptation with full covariance Gaussians. In: Proceedings of the INTERSPEECH (2006)
30. Vesely, K., et al.: Sequence-discriminative training of deep neural networks. In: Proceedings of the INTERSPEECH (2013)