Hindawi
Advances in Multimedia
Volume 2018, Article ID 8786308, 7 pages
https://doi.org/10.1155/2018/8786308

Research Article

Mobile Phone-Based Audio Announcement Detection and Recognition for People with Hearing Impairment

Yong Ruan ,[1,2] Yueliang Qian,[1] and Xiangdong Wang [1]

[1] *Beijing Key Laboratory of Mobile Computing and Pervasive Device, Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China*
[2] *University of Chinese Academy of Sciences, Beijing, China*

Correspondence should be addressed to Xiangdong Wang; xdwang@ict.ac.cn

Received 4 May 2018; Revised 27 July 2018; Accepted 5 August 2018; Published 16 August 2018

Academic Editor: Yong Luo

Automatic audio announcement systems are widely used in public places such as transportation vehicles and facilities, hospitals, and banks. However, these systems cannot be used by people with hearing impairment. That brings great inconvenience to their lives. In this paper, an approach of audio announcement detection and recognition for the hearing-impaired people based on the smart phone is proposed and a mobile phone application (app) is developed, taking the bank as a major applying scenario. Using the app, the users can sign up alerts for their numbers and then the system begins to detect audio announcements using the microphone on the smart phone. For each audio announcement detected, the speech within it is recognized and the text is displayed on the screen of the phone. When the number the user input is announced, alert will be given by vibration. For audio announcement detection, a method based on audio segment classification and postprocessing is proposed, which uses a SVM classifier trained on audio announcements and environment noise collected in banks. For announcement speech recognition, an ASR engine is developed using a GMM-HMM-based acoustic model and a finite state transducer (FST) based grammar. The acoustic model is trained on audio announcement speech collected in banks, and the grammar is human-defined according to the patterns used by the automatic audio announcement systems. Experimental results show that character error rates (CERs) around 5% can be achieved for the announcement speech, which shows feasibility of the proposed method and system.

1. Introduction

Voice is one of the most natural and important communication methods for human beings [1]. It is playing an increasingly important role and bringing great convenience to our daily lives. As an instance, automatic audio announcement systems are widely used in public places such as transportation vehicles and facilities, hospitals, and banks. For example, nowadays in China, each customer in a bank gets a number on his/her arrival and waits for his/her turn of service until the audio announcement containing the number tells him/her to which counter to go. However, these systems cannot be used by people with hearing impairment. Obviously, this brings great inconvenience to their lives. Although sometimes the audio announcements are companioned by text announcements displayed on screens, it is still inconvenient since the hearing-impaired people need to be always paying attention on the screens, which is tiring and may lead to missing of the information.

To make the announcements also available to hearing-impaired people, wireless sensor networks or short messenger systems are adopted to provide reminders via other methods instead of voice [2]. The disadvantage of these systems is that additional systems need to be deployed in the application scenarios, which is often difficult and expensive.

In this paper, we propose a mobile phone-based solution which avoids the deployment of additional systems in public places. An application (app) installed on the mobile phone can automatically detect and recognize audio announcements and remind the user by vibration and text on the mobile phone. There are two major challenges in the system: the detection of audio announcements in audio collected in real-world public places and recognition of the speech in the audio announcements detected.

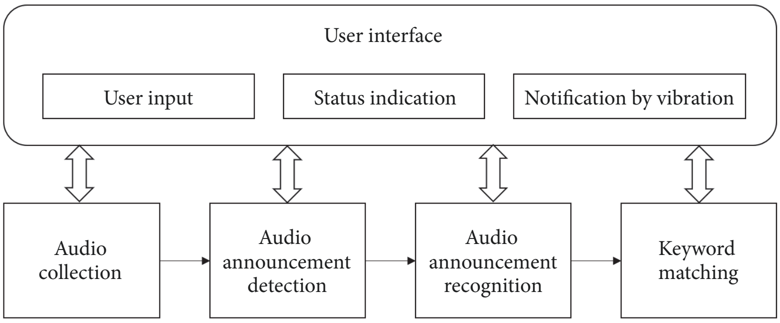

User interface
User input
Status indication
Notification by vibration
Audio collection
Audio announcement detection
Audio announcement recognition
Keyword matching

Figure 1: Framework of the proposed system.

The detection of the audio announcement can be viewed as a task of content-based audio classification, which is inherently a difficult problem of pattern recognition. There are two main issues for audio classification: the selection of audio features and the choice of classifiers [3]. For the selection of features, Pfeiffer et al. [4] proposed a theoretical framework that uses a series of perceptual features for automatically audio content analysis. Li et al. [5] studied the effect of a total of 143 audio features, showing that cepstrum-based features such as Mel-frequency cepstral coefficients (MFCCs) and Linear Predictive Cepstral Coefficients (LPCCs) are more effective than short-term and spectral features in audio classification. Feature-based fusion [6] and adaptive approach [7] are also useful in audio classification. As for classifiers used in audio classification [8], Lu et al. [9] used support vector machine (SVM) to classify audio into five categories: silence, music, background sound, the voice of a pure speaker, and the voice of a speaker under noise or music, which is similar to the Cuckoo algorithm [10].

Automatic speech recognition (ASR) techniques have been studied for decades and used in many real-world applications. Earlier ASR systems used Hidden Markov Model (HMM) for acoustic model. In recent years, the application of Deep Neural Networks (DNNs) [11] has significantly improved the accuracy of speech recognition. Neural networks such as Convolutional Neural Networks (CNNs) [12] and Recurrent Neural Networks (RNNs) [13, 14] are used and proved to be promising [15], and end-to-end speech recognition methods [16, 17] using Connectionist Temporal Classification (CTC) [18] and attention [19, 20] are also proposed to further improve the performance. Although the DNN-based approaches are reported to outperform HMM-based systems, very large training data are needed to train the DNNs. For applications on mobile phones, there are also many open cloud services of speech recognition. For example, in China, Baidu, iFLYTEK, Unisound, and other companies provide convenient remote interfaces of speech recognition which can be called by apps on mobile phones. However, these services mainly focus on speech of general domains and low-noise environments and may yield poor performance for speech in special scenarios such as far-field audio announcements in noisy public places [21, 22].

In this paper, an approach of audio announcement detection and recognition for the hearing-impaired people based on the smart phone is proposed and an Android app is developed, taking the bank as a major applying scenario. Using the app, the users can sign up alerts for their numbers and then the system begins to detect audio announcements using the microphone on the smart phone. For each audio announcement detected, the speech within it is recognized and the text is displayed on the screen of the phone. When the number the user input is announced, alert will be given by vibration. For audio announcement detection, a method based on audio segment classification and postprocessing is proposed, which uses a SVM classifier trained on audio announcements and environment noise collected in banks. For announcement speech recognition, an ASR engine is developed using a GMM-HMM-based acoustic model and a finite state transducer (FST) based grammar. The acoustic model is trained on audio announcement speech collected in banks, and the grammar is human-defined according to the patterns used by the automatic audio announcement systems. Experimental results show that character error rates (CERs) around 5% can be achieved for the announcement speech, which shows feasibility of the proposed method and system.

2. System Overview

In this paper, a method of audio announcement detection and recognition is proposed for the hearing-impaired people. Based on the method, we developed a mobile phone-based system, i.e., an Android app that can remind hearing-impaired people of audio announcements containing specified keywords, e.g., numbers. The current version of the app mainly focuses on the scenario of banks, although the proposed method can be used for almost all public places with audio announcements.

The framework of the proposed system is shown in Figure 1. The user interface receives user input (a keyword, e.g., a number) and starts monitoring of the keyword. During the monitoring, status information is displayed, and when the keyword is detected, the user interface will notify the user by vibration. The monitoring of the keyword is achieved by real-time detection and recognition of audio announcements in the ambient audio collected by the mobile phone. For each audio announcement detected, keyword matching is performed between the text of the audio announcement and

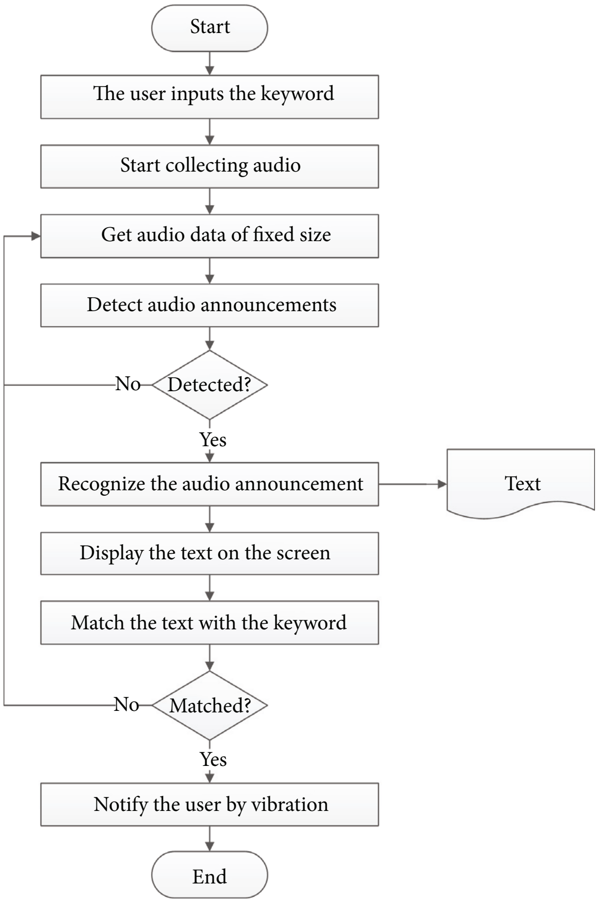
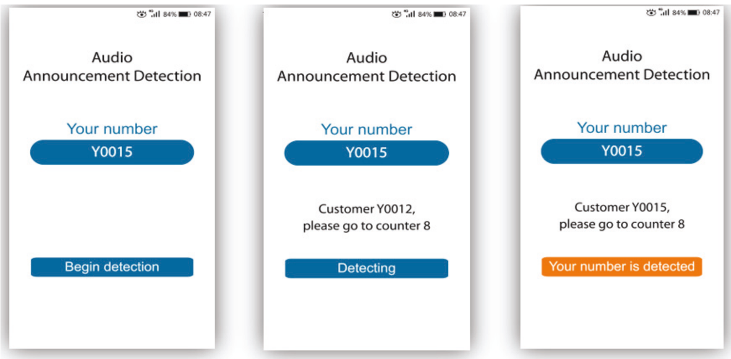

Start
The user inputs the keyword
Start collecting audio
Get audio data of fixed size
Detect audio announcements
Detected?
No
Yes
Recognize the audio announcement
Text
Display the text on the screen
Match the text with the keyword
Matched?
No
Yes
Notify the user by vibration
End

Figure 2: Flowchart of the proposed system.



Customer Y0012, please go to counter 8

Customer Y0015, please go to counter 8

(a) After launching the app, the user inputs the keyword/number (Y0015 in this case) and starts monitoring of the keyword/number by clicking the *start* (开始提醒) button.

(b) While monitoring the keyword/number, audio announcements are detected and recognized, and the text of each announcement is displayed (the announcement related to Y0012 in this case).

(c) For each announcement detected, matching between its text and the user-input keyword/number is performed. If the match succeeds, the app will notify the user by vibration and highlighted text on the screen.

Figure 3: Screen shot of the app.

the user input keyword to decide whether the keyword is detected.

A more detailed procedure of the proposed systems is shown in the flowchart in Figure 2 and illustrated by the screen shots of the app in Figure 3. After starting launching the app, the user sets the keyword to be monitored, which is usually the number they have been assigned when arriving at the bank, as shown in Figure 3(a). Then, the mobile phone starts to continuously collect audio data and store the data into the buffer pool. At the same time, the system begins detection of audio announcements using the data in the buffer. Each time, audio data of a fixed size is processed.

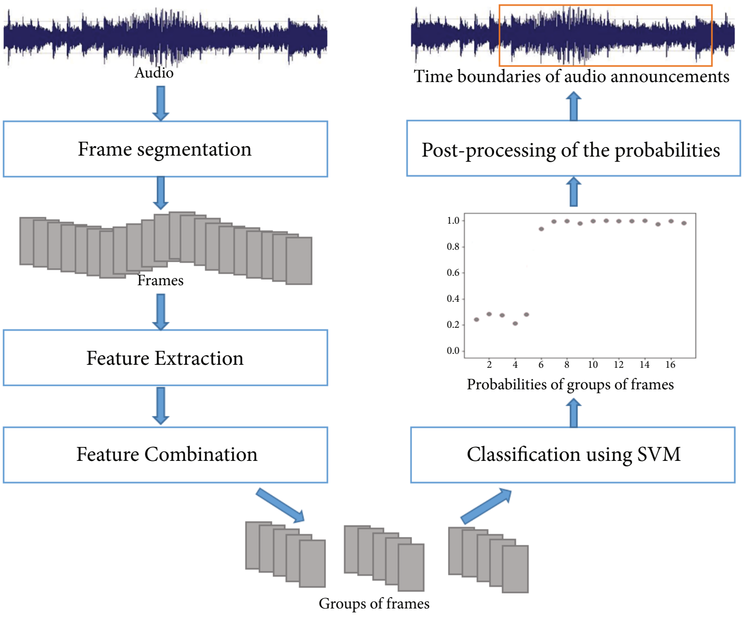


FIGURE 4: The procedure of audio announcement detection.

If audio announcements are detected in the data, the audio data containing announcement speech are input into the announcement recognition module, which is a speech recognition engine trained specially for audio announcements. The announcement recognition module transforms the audio into text and matches the text with the user-specified keyword. If the match succeeds, vibration will be triggered to notify the user, as shown in Figure 3(c). Considering that match failure may be caused by errors in speech recognition, during the monitoring, the recognition result of each audio announcement detected is also displayed on the mobile phone screen, as shown in Figure 3(b). With this auxiliary information, the user can understand current situation better and may correct speech recognition errors by themselves using context knowledge.

The core of the system is the automatic detection and recognition of audio announcements. Machine learning based methods are proposed for both the detection and recognition tasks, which will be detailed in the following sections. On the other hand, the methods for data collection and keyword matching are relatively simple and will not be further described.

3. Audio Announcement Detection

The procedure of audio announcement detection is as shown in Figure 4.

The audio data collected and stored in the buffer are first divided into frames with a length of 25 ms and without frame shift. Then preprocessing including pre-emphasis and Hamming windowing are performed and audio features are extracted for each frame. The features adopted are 13 MFCC coefficients, short time energy, and zero crossing rate. Therefore, a 15-dimension feature vector is extracted for each frame.

A classification-based scheme is used for audio announcement detection. Instead of classifying each frame, classification is performed for segments of 0.5 seconds, since segments can be more distinctive between audio announcement and background noise and segment-level classification is more efficient. Every 20 frames, which is related to an audio segment of 0.5 seconds, are combined into a group of frames, and the features of frames are combined into a 300-dimensional super-vector. The super-vector is then input into the classifier.

As for the classifier, Support Vector Machine (SVM) with RBF kernel is adopted due to its wide usage in content-based audio classification. An SVM model is trained with audio data collected in banks. Audio announcements are manually segmented from the audio and used as positive samples, and the remaining audio data without audio announcements are treated as negative samples. Since the audio announcements are relatively few, 1/5 of the negative samples are randomly selected and used for training. During the training stage, 8-fold cross-validation is performed to tune parameters.

By using the SVM to classify the audio for every 0.5 seconds, the probability of each 0.5-second segment being an audio announcement can be obtained. Then, postprocessing is performed to decide the starting and ending times of the audio announcement based on the probabilities. In our observation, we found that all the audio announcements in banks are within 6 seconds. Therefore, a sliding window of 6 seconds is adopted. For each time, 12 adjacent groups of frames are used for classification, and the 12 probabilities p_i, i=1, . . ., 12, are obtained. The average probability value p is then calculated as

$$p = \sum_{i=1}^{12} \frac{p_i}{12} \qquad (1)$$

If p is greater than a predetermined threshold, the 6-second segment is judged as an audio announcement.

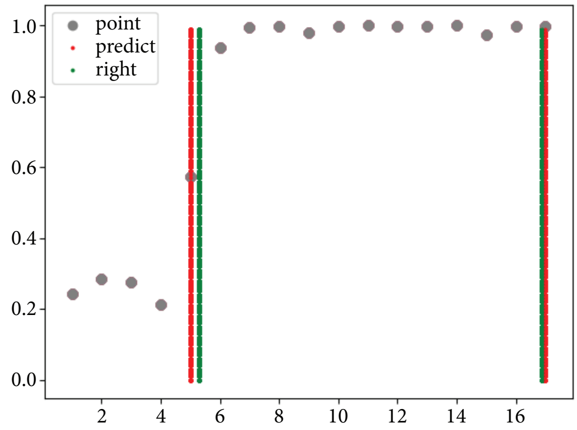


Figure 5: Example of postprocessing in audio announcement detection.

Figure 5 shows an example of the postprocessing. The abscissa indicates the time (second), and the ordinate indicates the probability value. Each point indicates the probability of the corresponding 0.5-second audio segment being an audio announcement. The red lines indicate the prediction of the starting and ending times of the audio announcement, and the green lines indicate the ground truth labelled by human.

4. Audio Announcement Recognition

After obtaining the audio announcements, the speech within them needs to be recognized to yield the text used for keyword matching. Due to the high-noise, and especially, the far collection distance of the audio collected in public places such as banks, general-purpose speech recognition systems which focus on speech collected by close talking microphones cannot be used. We have tried to use some cloud services of speech recognition. However, very poor recognition results were obtained (experimental results will be given in the following section), and some speech recognition engines even returned null for most of the speech. Therefore, we built a speech recognition engine for audio announcements in banks by ourselves, using the open source tool kit KALDI.

The challenge of building a speech recognition engine for audio announcements mainly lies in the lack of data. Therefore, we collected 27-hour audio data in 5 banks. The data contains 995 audio announcements in total. Although the amount is still small compared to other speech corpora, it should be noticed that the speech in the audio announcement is with limited vocabulary and of almost fixed pattern of expression. Therefore, a continuous speech recognition system with a small vocabulary and a simple grammar can be built using the data collected.

Due to the small amount of the training data, models based on deep neural network are not suitable. The traditional HMM-GMM model is adopted. During the training stage, the training data are divided into frames with 25 ms frame length and 10 ms frame shift. The 12-dimensional MFCC feature is extracted for each frame and, along with short time energy, forms a 13-dimensional vector. Then, after first-order and second-order differentiation, a 39-dimensional feature vector is finally obtained. There are 48 initials and vowels used in the dictionary as phonemes. Both mono-phone and triphone-based models are trained, and the training of the triphone model uses decision tree clustering for state binding to reduce the number of states. The Baum-Welch algorithm is used for training and beam search of Viterbi algorithm is used for decoding.

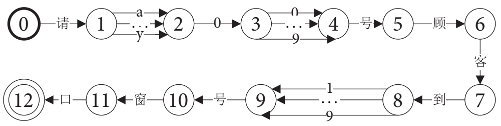


Figure 6: Finite state transducer of the grammar used.

For scenarios such as banks, hospitals, and transportation vehicles and facilities, audio announcements are mostly generated automatically and follow fixed patterns. For example, for banks in China, audio announcements are usually like "Customer A0017, please go to counter 8". Therefore, in our work, a grammar is defined according to the pattern of announcements in banks and used in the speech recognition system as language model. The grammar is represented by a finite state transducer, as shown in Figure 6, and is defined using the OpenFst tool in KALDI.

5. Experiments

5.1. Experiment Setup. The experimental data are audio data collected by mobile phones in 5 banks belonging to 3 different companies. The data are stored in 16 KHz, 16-bit, mono PCM WAV files. Information of the data is given in Table 1.

For audio announcement detection, the task is to detect audio announcements in the audio and give the time boundaries of the audio announcements. In the experiment, the input of this task is audio files collected in banks, and the output is the starting and ending times of the audio announcements in all the files. There may be multiple or no audio announcements in an audio file. The training set is formed by all audio in the first 4 banks and the test set contains audio in the fifth bank. The test data consists of 50 audio files with announcements and 10 files without audio announcements. The starting and ending times of the audio announcements in test data are labelled manually as the ground truth. If the deviation between the ground truth and the detected time is less than a given threshold, the detection is considered to be correct. We compute the recall rate, precision rate, and the F1 value to evaluate the overall detection performance.

For speech recognition, to analyse the effect of different data collection sites, two experiments are conducted by using different datasets. The first training set consists of 770 audio files, each containing one audio announcement. These audios are collected in four different banks, while the test data contains 132 audio files collected in a different bank, namely, the fifth one. The second training set consists of 802 audio files collected in all the five banks, and the second test

Table 1: Information of experimental data.

Bank	Duration (hours)	Number of announcements
Zhichunlu	10	276
Shuangyushu	4	97
Kexueyuan	7	426
Xili	4	157
Haidian	2	39
Total	27	995

Table 2: Experimental results of audio announcement detection.

	Precision rate	Recall rate	F1
Tolerance = 0.5s	0.822	0.740	0.778
Tolerance = 1.0s	0.855	0.940	0.895

Table 3: Results of audio announcement recognition.

	The first dataset (training data from 4 banks and test data from the fifth one)		The second dataset (training and test data both from 5 banks)	
	CER (%)	SER (%)	CER (%)	SER (%)
Mono-phone	4.11	34.09	5.12	33.33
Tri-phone	5.41	40.15	5.41	40.40
Baidu	85.54	100.00	73.79	100.00
iFLYTEK	54.67	98.34	38.86	95.41
Unisound	95.10	100.00	88.93	100.00

data contains 100 audio files collected also in the same five banks. The training data and test data do not contain same audio, even audio collected in the same day in one bank. The character error rate (CER) and the sentence error rate (SER) are used to measure the accuracy of speech recognition. To demonstrate the infeasibility of current general-purpose speech recognition engines, three cloud services of Mandarin speech recognition, namely, Baidu, iFLYTEK, and Unisound, are also used to recognize both test sets.

5.2. Experimental Results. For audio announcement detection, experimental results are shown in Table 2, where *tolerance* is the threshold that both the starting and ending time deviation should be less than. It can be seen that an F1 value of 0.895 and a high recall rate (94%) can be achieved for the 1-second tolerance. In fact, the speech recognition module does not require accurate time boundaries of the speech and can deal with speech with background noise. Therefore, the 1-second tolerance is enough for the speech recognition. An advantage of the proposed method is its high recall rate, since in the application scenarios, miss of the audio announcement will make the system useless while the user can to a degree tolerate some false alarms.

The experimental results of audio announcement recognition are shown in Table 3. It can be seen that the mono-phone model outperforms the tri-phone model. This may be due to the small amount of the training data. As for the two different data sets, the performances do not differ much and the CER on the first data set is even lower, which means the system is robust enough to be used in banks in which no training data is collected. For the mono-phone model, a CER about 5% is achieved, which shows feasibility of the proposed method and system. As for the general-purpose speech recognition services, the CERs are very high and the SERs are near 100%. This is because the engines mainly focus on speech data collected by close talking microphones and can not deal with far-field speech collected by a single mobile phone microphone.

6. Conclusions

This article describes a novel system that runs on a mobile phone. The system can be used by people with hearing impairment to avoid missing audio announcements they concern in public places. An approach of audio announcement detection and recognition is proposed and an Android app is developed, taking the bank as a major applying scenario. For audio announcement detection, a method based on audio segment classification and postprocessing is proposed, which uses a SVM classifier trained on audio announcements and environment noise collected in banks. For announcement speech recognition, an ASR engine is developed using a GMM-HMM-based acoustic model and an FST based grammar. Experimental results show that character error rates (CERs) around 5% can be achieved for the announcement speech, which shows feasibility of the proposed method and system. Future work includes extending the system usage to more public places and improvement of the keyword match module.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work is partly supported by Beijing Natural Science Foundation (4172058).

References

[1] B.-H. Juang and S. Furui, “Automatic recognition and understanding of spoken language - A first step toward natural human-machine communication,” *Proceedings of the IEEE*, vol. 88, no. 8, pp. 1142–1165, 2000.

[2] D. Chander and M. V. Sireesha, *Passenger bus alert system for easy navigation of blind*, 2004.

[3] L. Lu, H.-J. Zhang, and H. Jiang, “Content analysis for audio classification and segmentation,” *IEEE Transactions on Audio, Speech and Language Processing*, vol. 10, no. 7, pp. 504–516, 2002.

[4] S. Pfeiffer, S. Fischer, and W. Effelsberg, “Automatic audio content analysis,” in *Proceedings of the 1996 4th ACM International Multimedia Conference*, pp. 21–30, November 1996.

[5] D. Li, I. K. Sethi, N. Dimitrova, and T. McGee, “Classification of general audio data for content-based retrieval,” *Pattern Recognition Letters*, vol. 22, no. 5, pp. 533–544, 2001.

[6] H. K. Palo, M. N. Mohanty, and M. Chandra, “Efficient feature combination techniques for emotional speech classification,” *International Journal of Speech Technology*, vol. 19, no. 1, pp. 135–150, 2016.

[7] K. Khaldi, A.-O. Boudraa, and M. Turki, “Voiced/unvoiced speech classification-based adaptive filtering of decomposed empirical modes for speech enhancement,” *IET Signal Processing*, vol. 10, no. 1, pp. 69–80, 2016.

[8] S. Baghel, S. R. M. Prasanna, and P. Guha, “Classification of multi speaker shouted speech and single speaker normal speech,” in *Proceedings of the 2017 IEEE Region 10 Conference, TENCON 2017*, pp. 2388–2392, Malaysia, November 2017.

[9] L. Lu, H.-J. Zhang, and S. Z. Li, “Content-based audio classification and segmentation by using support vector machines,” *Multimedia Systems*, vol. 8, no. 6, pp. 482–492, 2003.

[10] W. Shi and X. Fan, “Speech classification based on cuckoo algorithm and support vector machines,” in *Proceedings of the 2nd IEEE International Conference on Computational Intelligence and Applications, ICCIA 2017*, pp. 98–102, China, September 2017.

[11] D. Yu and L. Deng, *Automatic speech recognition*, Springer london limited, 2016.

[12] D. Palaz, M. Magimai-Doss, and R. Collobert, “Analysis of CNN-based speech recognition system using raw speech as input,” in *Proceedings of the 16th Annual Conference of the International Speech Communication Association, INTERSPEECH 2015*, pp. 11–15, Germany, September 2015.

[13] H. Sak, A. Senior, K. Rao et al., “Learning acoustic frame labeling for speech recognition with recurrent neural networks,” in *Proceedings of the 40th IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2015*, pp. 4280–4284, Australia, April 2014.

[14] H. Soltau, H. Liao, and H. Sak, “Neural Speech Recognizer: Acoustic-to-Word LSTM Model for Large Vocabulary Speech Recognition,” in *Proceedings of the Interspeech 2017*, pp. 3707–3711.

[15] J. Tebelskis, *Speech recognition using neural networks*, Siemens AG, 1995.

[16] D. Amodei, S. Ananthanarayanan, and R. Anubhai, “Deep speech 2: End-to-end speech recognition in english and mandarin,” in *Proceedings of the International Conference on Machine Learning*, pp. 173–182, 2016.

[17] D. Bahdanau, J. Chorowski, D. Serdyuk, P. Brakel, and Y. Bengio, “End-to-end attention-based large vocabulary speech recognition,” in *Proceedings of the 41st IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2016*, pp. 4945–4949, China, March 2016.

[18] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, “Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks,” in *Proceedings of the 23rd International Conference on Machine Learning, ICML 2006*, pp. 369–376, Pittsburgh, Pa, USA, June 2006.

[19] D. Bahdanau, C. Kyunghyun, and Y. Bengio, “Neural Machine Translation by Jointly Learning to Align and Translate,” in *In International Conference on Learning Representa-tions*, 2015.

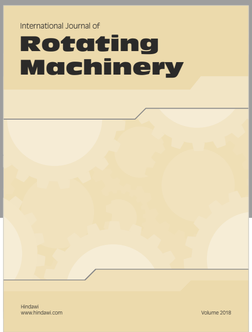
[20] W. Chan, N. Jaitly, Q. Le, and O. Vinyals, “Listen, attend and spell: A neural network for large vocabulary conversational speech recognition,” in *Proceedings of the 41st IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2016*, pp. 4960–4964, China, March 2016.

[21] N. Morgan and H. Bourlard, “Continuous speech recognition,” *IEEE Signal Processing Magazine*, vol. 12, no. 3, pp. 24–42, 1995.

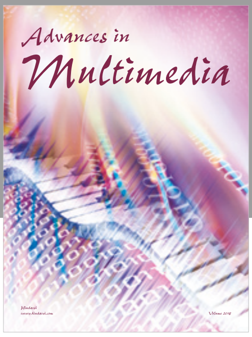
[22] B. M. J. Leiner, *Noise-Robust Speech Recognition*, 2003.

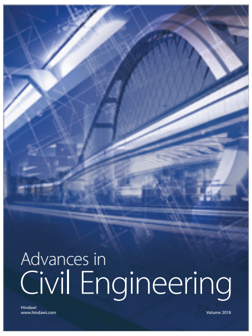
Journal of
Engineering

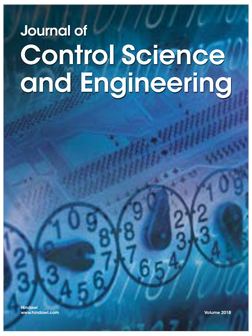
The Scientific World Journal


International Journal of
Rotating Machinery

Journal of
Sensors

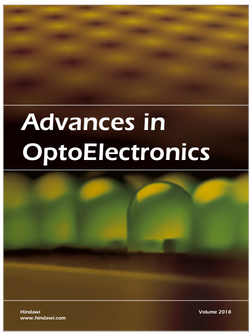

Advances in
Multimedia


Advances in
Civil Engineering


Journal of
Control Science and Engineering

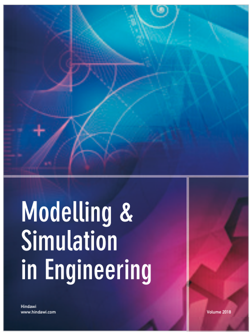
Journal of
Robotics

Journal of
Electrical and Computer Engineering

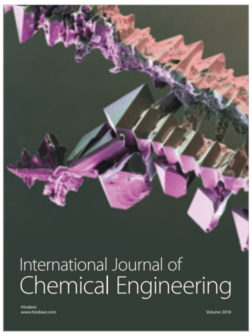

Advances in
OptoElectronics

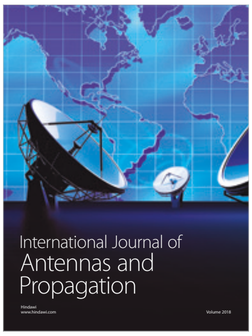
VLSI Design

International Journal of
Navigation and Observation

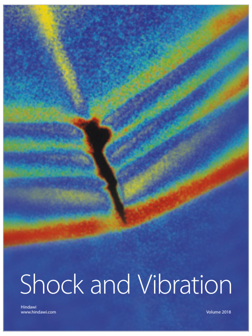

Modelling & Simulation in Engineering

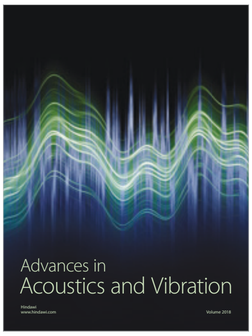
International Journal of
Aerospace Engineering


International Journal of
Chemical Engineering


International Journal of
Antennas and Propagation

Active and Passive
Electronic Components


Shock and Vibration


Advances in
Acoustics and Vibration