Cross-Covariance-Based Features for Speech Classification in Film Audio

Matt Benatan and Kia Ng
University of Leeds,
School of Computing,
Leeds LS2 9JT, United Kingdom
mattbenatan@gmail.com, k.c.ng@leeds.ac.uk

***Abstract*— As multimedia becomes the dominant form of entertainment through an ever increasing range of digital formats, there has been a growing interest in obtaining information from entertainment media. Speech is one of the core resources in multimedia, providing a foundation for the extraction of semantic information. Thus, detecting speech is a critical first step for speech-based information retrieval systems. This work focuses on speech detection in one of the dominant forms of entertainment media: feature films. A novel approach for voice activity detection (VAD) in film audio is proposed. The approach uses correlation to analyze associations of Mel Frequency Cepstral Coefficient (MFCC) pairs in speech and non-speech data. This information then drives feature selection for the creation of MFCC cross-covariance feature vectors (MFCC-CCs) which are used to train a random forest classifier to solve a binary speech/non-speech classification problem on audio data from entertainment media. The classifier performance is evaluated on a number of test sets and achieves a classification accuracy of up to 94%. The approach is also compared with state of the art and contemporary VAD algorithms, and demonstrates competitive results.**

Keywords- voice activity detection; speech detection; binary classification; film audio; entertainment media

I. INTRODUCTION

Consumption of multimedia has become ubiquitous, with TV, films, games, and digital music now providing the majority of our entertainment in a range of easily accessible formats. With this rise in multimedia, there is a continually increasing interest in obtaining information from media - using it to understand human interaction and behavior [1], and to extract semantic information that can be used in the creation of metadata [2]. Speech classification plays a key role in data extraction through detecting speech regions in audio or video data. These regions can then be used for further feature extraction, e.g. speech recognition. While many speech detection techniques exist, few have been developed specifically for use with one of our most challenging and popular forms of media: film. Unlike radio and news broadcasts, films contain an extremely diverse range of speech and other audio content. Film introduces challenges that are not present in most natural scenarios, such as speech in the presence of highly dynamic background noise and sound effects, or heavily manipulated speech, where sound design has been used to create unnatural voice characteristics through the addition of harmonics and distortion.

We present a novel approach for speech detection developed specifically for classification of speech within film audio. This approach aims to account for unusual voice characteristics by analyzing the relationships between pairs of spectral features within speech and non-speech data. We use the process to identify Mel Frequency Cepstral Coefficient (MFCC) pairs which are then processed to create cross-covariance-based feature vectors (MFCC-CCs). MFCC covariance statistics have been used previously for audio classification tasks, such as in [3] and [4], where covariance is used alongside other statistical representations of MFCC data, resulting in as many as 60 dimensions per frame (as described in [4]). In this work, cross-correlation is used to select specific MFCC pairs which demonstrate the greatest difference in correlation between speech and non-speech data. Cross-covariance vectors for the five highest scoring MFCC pairs are then created, providing a single vector which represents the covariance relationship for each pair. The resulting feature vector is comprised of five speech-sensitive MFCC-CC features per frame, thus reducing dimensionality from 13 MFCCs to five MFCC-CC features. Through using this feature vector with a random forest classifier, we have achieved a classification accuracy of 94% on challenging audio data.

II. BACKGROUND

Recent developments in mixed-audio speech detection have demonstrated high accuracy [5], however, while using mixed audio signals, the datasets used in much of the work to date is still fairly constrained. These include radio broadcasts [6], news broadcasts [7], and speech detection in the presence of background noise [8]. Speech detection in these scenarios is likely to be a simpler task than speech detection within film audio. This is due to the dynamic nature of film audio: not only does it contain various types of background noise, but the acoustic environments change frequently (simulated or

This work is supported by the UK Engineering and Physical Sciences Research Council (EPSRC) and ZOO Digital Group plc.
DOI reference number: 10.18293/DMS2015-023

otherwise, e.g. via reverb effects [9]) and the format makes use of many synthetic sound effects [10], which can obscure speech information in the audio. As well as this, voice synthesis or distortion is now also common in feature films [10], all of which make speech detection more challenging when using typical spectral features. To address this, we have developed an approach for speech detection that uses cross-covariance to represent the relationship between pairs of MFCCs [11]. This reduces feature dimensionality, resulting in a feature set designed to improve speech/non-speech discrimination. The resulting feature vector is used to train a learning machine to perform binary classification (speech/non-speech) on an annotated ground-truth dataset. Results demonstrate an accuracy of between 86.15% and 87.26%, which are promising performance statistics when considering the challenging nature of the dataset.

An approach discussed in [6], for classifying speech/non-speech in radio broadcasts, exploits spectro-temporal variations of speech signals via short time Fourier transforms (STFTs) to discriminate between speech and non-speech signals. This has demonstrated good performance on their data set, however, this approach applies a median filter or approximately 10 seconds duration to the classifier output. Thus, it is primarily useful for broadly classifying sections of audio, rather than for higher resolution speech activity detection. Furthermore, the data used is sourced exclusively from radio broadcasts, and is thus not reflective of film audio content, likely being less dynamic and thus simplifying the classification problem.

Another recent approach described in [5] uses a voice activity detector based on Long Short-Term Memory Recurrent Neural Networks (LSTM-RNN). This demonstrates good performance on a synthetic test validation set, with an average equal error rate (EER) of 10.4%, outperforming the state-of-the-art SOHN algorithm. However, it is less effective on film audio, with an average EER of 33.2%.

One film-centered approach [12] utilizes bilingual audio streams for speech detection. This identifies speech segments through correlating spectral coefficients between two different language tracks, and demonstrates an accuracy of between 84% and 87% in classifying clean and noisy speech in film audio. While this approach demonstrates good performance on film data, it requires bilingual audio tracks to perform classification, and thus would not work with single language audio data.

Another approach discussed in [13] uses a dataset comprised entirely of television material (thus similar to film) and looks to differentiate between speech and music data. This uses discrete wavelet transforms (DWT) as the audio feature and performs classification via a support vector machine. While this performs with an accuracy of up to 94.5%, the approach is focused on discriminating between speech or music data, and thus does not consider environmental noise, silence, sound effects and other sonic components common to film audio.

Several other reviewed approaches have demonstrated an accuracy of >90%, however, these either have limited data, such as [14], which has only 9 main speakers in its dataset, or make use of non-film audio, such as [7], whose data includes radio and news broadcasts (which typically do not have the same sonic variance as film data).

III. Proposed Approach

A. Process Overview

The speech classification process consists of three core stages. The first of these is feature selection, which analyzes the audio data using cross-correlation to determine which features yield the most useful information to discriminate between speech and non-speech data. The second stage consists of processing this information to create the MFCC-CC feature vectors, and in the third stage a classifier is fit to a training set of ground-truth labelled data.

B. Feature Selection

Numerous approaches for spectral feature parameterization exist [15], however MFCCs are one of the most frequently used spectral features in both automatic speech recognition (ASR) [16] and voice activity detection (VAD) [17]. Given their wide adoption in speech processing, MFCCs have been chosen as the method of representing spectral features in this work. Within this application we replace the zeroth MFCC with the log of the total frame energy, as this has proven to be useful in speech processing applications [18] [19].

Feature selection is achieved using cross-correlation to analyze the difference in cross-MFCC similarities in speech and non-speech data from the training set. A correlation coefficient is obtained for each MFCC with respect to each of the other MFCCs. This is done separately for speech and non-speech data. The speech/non-speech difference in the resulting correlation coefficients for each MFCC feature pair is obtained. This is used to determine which feature pairs demonstrate the greatest change in correlation between speech and non-speech data. The Pearson product-moment correlation coefficient, ρ, is obtained from the covariance matrix (C) of a pair of MFCC feature vectors via the coefficient matrix P_{ij}:

$$P_{ij} = \frac{c_{ij}}{\sqrt{c_{ii} * c_{jj}}} \quad (1)$$

The correlation coefficient has a value between -1 and 1, where 1 denotes total positive correlation, and -1 denotes total negative correlation.

The MFCC pairs are chosen based on the difference between their speech and non-speech correlation coefficients. Figure 1 shows the resulting correlation coefficient differences. Higher values indicate a greater variance in the MFCC pair relationships between speech and non-speech data. This in turn indicates that the pairs are more likely to provide information relating to the presence/absence of speech spectral data, thus facilitating more effective speech/non-speech discrimination.

MFCC	0	1	2	3	4	5	6	7	8	9	10	11	12
0	0	0.04	0.1	0.21	0.18	0.06	0.34	0	0.06	0.23	0.12	0.21	0.1
1	0.04	0	0.1	0.01	0.31	0.06	0.11	0.02	0.2	0.02	0.19	0.08	0.15
2	0.1	0.1	0	0.06	0.13	0.08	0.08	0	0.24	0.12	0.15	0.13	0.04
3	0.21	0.01	0.06	0	0.12	0.15	0.04	0.13	0.04	0.03	0.1	0.07	0.09
4	0.18	0.31	0.13	0.12	0	0.04	0.07	0.04	0.04	0.01	0.03	0.01	0.11
5	0.06	0.06	0.08	0.15	0.04	0	0.1	0.12	0.03	0.04	0.08	0.06	0.02
6	0.34	0.11	0.08	0.04	0.07	0.1	0	0.2	0.03	0.1	0.07	0.1	0.2
7	0	0.02	0	0.13	0.04	0.12	0.2	0	0.07	0.13	0.07	0.17	0.17
8	0.06	0.2	0.24	0.04	0.04	0.03	0.03	0.07	0	0.1	0.19	0.05	0.18
9	0.23	0.02	0.12	0.03	0.01	0.04	0.1	0.13	0.1	0	0.07	0.15	0.03
10	0.12	0.19	0.15	0.1	0.03	0.08	0.07	0.07	0.19	0.07	0	0.12	0.2
11	0.21	0.08	0.13	0.07	0.01	0.06	0.1	0.17	0.05	0.15	0.12	0	0.17
12	0.1	0.15	0.04	0.09	0.11	0.02	0.2	0.17	0.18	0.03	0.2	0.17	0

Figure 1. Matrix of MFCC pair correlation coefficient differences between speech and non-speech data. Darker squares indicate greater values.

C. Feature Vector Processing

The final MFCC-cross-covariance feature vectors are attained by computing the cross-covariance of the MFCC pairs corresponding to the top *n* correlation coefficient differences. For each pair of MFCCs, the cross-covariance vector is obtained through computing the cross-covariance of segments of the two signals along their length via a rectangular sliding window:

$$(f * g)_i \stackrel{\text{def}}{=} \sum_j f_j * g_{i+j} \tag{2}$$

$$f = v1_{k:k+w}, g = v2_{k:k+w} \; for \; all \; k \; in \; v1, v2$$

where *v1* is the first MFCC vector, *v2* is the second MFCC vector, *k* is the index and *w* is the size of the sliding window.

As temporal information has proven to be useful in speech classification problems [5, 6], a window size of 450ms has been used for *w*. This was determined based on average phoneme duration being around 176ms [20]. As such, a frame size of 450ms is therefore long enough to account for multiple phonemes, thus avoiding false classification of brief speech-like phenomena, but still allowing for the detection of finer resolution (sub-1s duration) speech features.

D. Classification and Tuning

Classification is achieved through the use of a learning machine trained on the MFCC-cross-covariance (MFCC-CC) features from the annotated training data set. In this case, random forests were chosen as the classifier based on their strong performance in speech classification applications [6, 21]. The random forest classifier was investigated using varying numbers of MFCC-CC features and a range of estimators (trees per forest) in order to determine optimal parameters for classification.

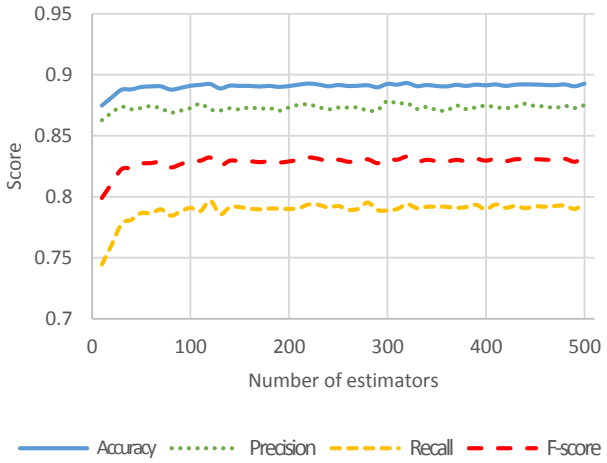

Figure 2. Random forest classification results using a range of estimators.

While testing numbers of estimators, it was found that the performance metrics stabilize after approximately 150 estimators (Figure 2), with little-to-no performance advantage achieved after this. Furthermore, previous work on random forest-based speech classifiers has demonstrated that optimal performance is achieved with the use of 200 estimators [6]. As such, the number of estimators for the random forest classifier was set at 200.

To test the impact of the number of MFCC-CC vectors used, vectors were added in order of significance, with the most significant relating to the MFCC pair with the greatest correlation coefficient difference across speech and non-speech data. Results from this (Figure 3) demonstrate that classification performance improves dramatically up to three features, and stabilizes at around five features. Therefore, five features were chosen as the optimal setting, as there was negligible gain in performance after this point.

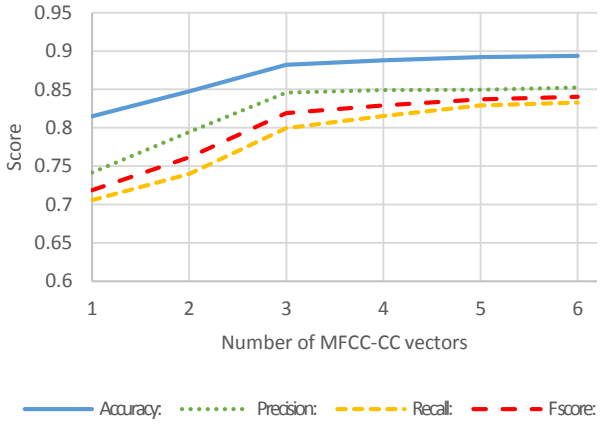

Figure 3. Random forest classification results with escalating numbers of MFCC-CC vectors.

IV. Case Study Design

Unlike other forms of data used within voice activity detection tasks, such as speech recordings in various acoustic environments [8] or synthesized acoustic environments, film is unique in that it is intentionally mixed [9]. While it may be intuitive to assume that this would make speech detection a simpler task (as the speech is mixed to be intelligible), this has proven not to be the case when testing a number of state-of-the-art voice activity detection algorithms on data from feature films [5]. This suggests that the intentional mixing of film audio separates it from audio data used in other typical VAD scenarios. As such, we have focused solely on the use of audio data from film – ensuring that both the training and test sets use intentionally mixed audio.

Two test scenarios have been used. The first uses a data set consisting of 120 minutes of data taken from four 30 minute segments of four feature films. To maximize usefulness, a cross-validation approach is used, whereby the data is reconfigured four times for each test. Each iteration uses 90 minutes of data for the training set (from three films), and 30 minutes of data for the test set (from the remaining film). This ensures that the classifier is naïve to the test data and maximizes testing cycles for the validation test set. The second test scenario uses all 120 minutes of data from the cross-validation set for training, and uses the films detailed in [5] as test data. This has been done to provide a direct comparison between the MFCC-CC approach, the approach from [6] and the results described in [5] (which includes results from testing the VAD described in [22] on feature film data).

The data has been manually annotated to provide a human-defined ground-truth, whereby sections are labeled as either speech or non-speech. The non-speech content consists of various audio mixtures including: silence, traffic noise, crowd noise, gunfire, engine noise, music (with and without singing), and other synthetic sound effects and sound design components. The speech content contains a number of speech varieties, including: speaking (various volumes), whispering, and shouting. Speech content is also mixed with the range of background audio (similar to that described for the non-speech content). The degree of variation in both speech and non-speech samples is pseudo-random according to individual film content.

V. Results

A. Initial Testing Results

Initial testing indicated strong performance of the MFCC-CC classifier, with an average accuracy of 89.2% (see Table I). Strong performance was also observed when testing on an animated feature film using training data from non-animated content. This indicates that the approach is capable of handling a-typical speech characteristics, as the animated content contains a significant amount of extreme/accentuated voice characteristics, for example the voice of the *Gingerbread Man* character in *Shrek 3*.

TABLE I. Classification results from random forest classifier trained on MFCC-CC features

Test set* | *Genre	***Accuracy***	***Precision***	***Recall***	***Fscore***
Constantine* | *action/horror	0.903	0.902	0.794	0.844
Shrek 3* | *animated/fantasy	0.861	0.792	0.789	0.790
Knocked Up* | *romantic comedy	0.881	0.783	0.889	0.833
Blood Diamond* | *drama/thriller	0.924	0.920	0.845	0.881
Mean	0.892	0.849	0.829	0.837

The MFCC-CC classifier was also evaluated using receiver operatic characteristics (ROC), a common method of assessing binary classifier performance. The ROC curves in Figure 4 indicate strong performance, with an average area under curve (AUC) of 0.955 (see Table II), indicating that the classifier exhibits strong discrimination between the two classes. The equal error rate (EER) observed here further indicates strong system accuracy, with an average EER of 11.1% achieved across the four test scenarios. This suggests better performance than the VAD in [5], which achieved an average EER of 33.2% on film audio data.

To assess performance with respect to [6], an implementation of the classifier used by Sonnleitner *et al.* was trained and tested using the cross-validation approach described in section IV. In [6] a median filter is used on the classification output. To assess equivalent performance, the median filter is not applied here, as a median filter has not been used on the MFCC-CC classifier output. Thus, only the raw classifier output is considered.

TABLE II. AUC and EER from receiver operating characteristic plot

	Const.	***Shrek 3***	***Kno. Up***	***Bl. D.***	***Mean***
AUC	0.969	0.925	0.954	0.973	0.955
EER [%]	9.5	15.0	11.6	8.1	11.1

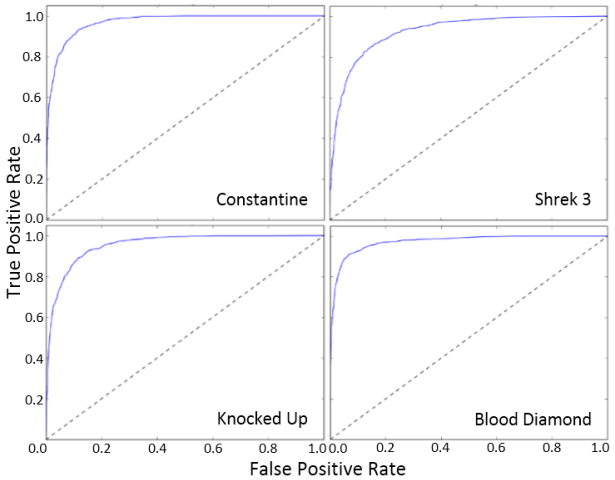

Figure 4. Receiver operatic characteristic curves for MFCC-CC classification results from initial testing.

TABLE III. CLASSIFICATION RESULTS FROM RANDOM FOREST CLASSIFIER TRAINED ON FEATURES DESCRIBED IN [6]

Test set | genre	*Accuracy*	*Precision*	*Recall*	*Fscore*
Constantine | action/horror	0.714	0.642	0.315	0.423
Shrek 3 | animated/fantasy	0.701	0.642	0.228	0.337
Knocked Up | romantic comedy	0.678	0.539	0.224	0.317
Blood Diamond | drama/thriller	0.701	0.637	0.236	0.344
Mean	0.699	0.615	0.251	0.355

As demonstrated when comparing Table I and Table III, the MFCC-CC approach achieves greater results across all performance statistics used for evaluation, thus early investigations indicated that the proposed MFCC-CC features are more effective for speech classification when compared to the feature proposed in [6].

B. Further Testing Results

Further investigations applied the MFCC-CC approach to whole feature films in order to provide a more comprehensive evaluation of its performance with respect to existing methods. The methods used for comparison were a long-standing state of the art VAD approach used to provide baseline performance statistics [22], as well as approaches that have demonstrated strong performance on entertainment media [5][6].

Results in Table IV indicate that the approach from [6] demonstrated competitive performance against both [5] and [22], however the MFCC-CC approach exceeds the performance of all methods investigated, with greater AUC values for all test sets and lower EER.

TABLE IV. COMPARISON OF VAD APPROACHES

	AUC			
Test set	***[5]***	***[22]***	***[6]***	***MFCC-CC***
I Am Legend	0.704	0.567	0.718	0.921
Kill Bill Vol. 1	0.627	0.554	0.800	0.893
Saving Private Ryan	0.743	0.577	0.717	0.946
The Bourne Identity	0.685	0.603	0.730	0.977
Mean	0.690	0.575	0.741	0.934
[%]	**EER**			
ALL	33.18	45.73	31.41	13.49

TABLE V. PERFORMANCE STATISTICS OF MFCC-CC APPROACH AND CLASSIFIER FROM [6] WHEN APPLIED TO WHOLE-FEATURE-FILM DATA SET

Test set	*Accuracy*		*Precision*		*Recall*		*Fscore*	
IAL	**0.88**	0.81	**0.62**	0.47	**0.81**	0.17	**0.70**	0.25
KB.1	**0.84**	0.79	**0.64**	0.62	**0.72**	0.26	**0.68**	0.37
SPR	**0.87**	0.77	**0.91**	0.45	**0.66**	0.29	**0.77**	0.35
TBI	**0.94**	0.76	**0.88**	0.45	**0.88**	0.25	**0.87**	0.32
Mean	**0.88**	0.78	**0.76**	0.50	**0.77**	0.24	**0.75**	0.32

Left columns (bold): MFCC-CC results. Right columns: results from approach described in [6]

Table V provides a more detailed performance comparison of the MFCC-CC approach and [6] (as this demonstrated the most competitive results in Table IV). The MFCC-CC approach demonstrates some reduced performance when compared to the initial testing results in Table I, however, this was anticipated given the limited training set and larger test set. Despite this, the approach continues to exhibit competitive results, outperforming [6] across all performance metrics. In particular, it can be seen that while the approach from [6] demonstrates relatively strong accuracy scores, significantly greater F-score values for our approach can be observed, indicating more robust performance.

VI. CONCLUSIONS AND FUTURE DIRECTION

The results presented here demonstrate strong performance of the proposed MFCC-CC speech detection approach, yielding performance metrics which exceed those of state of the art and other contemporary VAD approaches applied to feature film audio data. While these results are encouraging, more comprehensive testing is underway in order to gain further insight into the performance of the proposed approach on a larger data set. Given the small size of the training set used here (120 minutes), it will be particularly interesting to investigate the effects of more training data on classification performance, and to explore VAD performance on a greater variety of film genres and across multiple languages.

Further investigations will also explore the use of MFCC-CC features with other classifiers, such as support vector machines, and will examine the possibility of expanding the feature selection method to explore whether genre-specific MFCC feature pairs can be utilized to enhance classifier performance. The long-term goal of this work is to apply audio speech detection in combination with visual features to gain a better understanding of their associations and to develop automated solutions for film post-production workflows.

REFERENCES

[1] M. Benatan and K. Ng, "Multimodal Feature Matching for Event Synchronization," in Proceedings of the 19th International Conference on Distributed Multimedia Systems, Brighton, UK, 2013.

[2] Z. Rasheed, Y. Sheikh and M. Shah, "On the Use of Computable Features for Film Classification," in IEEE Transactions on Circuits and Systems for Video Technology, vol. 15, 2005, pp. 52-64.

[3] J. Bergstra, M. Mandel and D. Eck, "Scalable Genre and Tag Prediction with Spectral Covariance," in Proceedings of the 11th International Society for Music Information Retrieval Conference, Utrecht, Netherlands, 2010.

[4] D. P. W. Ellis, X. Zeng and J. H. McDermott, "Classifying Soundtracks with Audio Texture Features," in the 36th International Conference on Acoustics, Speech and Signal Processing, Prague, Czech Repbulic, 2011.

[5] F. Eyben, F. Weninger, S. Squartini and B. Schuller, "Real-life Voice Activity Detection with LSTM Recurrent Neural Networks and an Application to Hollywood Movies," in Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, Vancouver, BC, USA, May 2013, pp. 483-487.

[6] R. Sonnleitner, B. Niedermayer, G. Widmer and J. Schlüter, "A Simple and Effective Spectral Feature for Speech Detection in Mixed Audio Signals," in Porceedings of the 15th International Conference on Digital Audio Effects, York, UK, 2012.

[7] L. Lu, H-J Zhang and H. Jiang, "Content Analysis for Audio Classification and Segmentation," in IEEE Transactions on Speech and Audio Processing, vol. 10, October, 2002, pp. 504-516.

[8] J. Bach, J. Anemueller and B. Kollmeier, "Robust Speech Detection in Real Acoustic Backgrounds with Perceptually Motivated Features," in Speech Communication, vol. III, May, 2011, pp. 690-706.

[9] T. Holman, Sound for Film and Television. Kidlington, Oxford: Elsevier, 2010.

[10] R. Viers, The Sound Effects Bible. Studio City, CA: Michael Wiese Productions, 2008.

[11] S. B. Davis and P. Mermelstein, "Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences," in IEEE Transactions on Acoustic, Speech and Signal Processing, vol. 28, 1980, pp. 357-366.

[12] A. Tsiartas, P. Ghosh, P.G. Georgiou and S. Narayanan, "Bilingual Audio-subtitle Extraction Using Automatic Segmentation of Movie Audio," in Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, Prague, Czech Republic, 2011.

[13] T. Ramalingam and P. Dhanalakshmi, "Speech/Music Classification Using Wavelet Based Geature Extraction Techniques," Journal of Computer Science, vol. 10, 2014, pp. 34-44.

[14] J. Pinquier and C. Senac, "Speech and Music Classification in Audio Documents," in Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, Orlando, FL, USA, 2002.

[15] T. Drugman and Y. Stylianou, "Fast Inter-Harmonic Reconstruction for Spectral Envelope Estimation in High-Pitched Voices," in IEEE Signal Processing Letters, vol. 21, July, 2014, pp. 1418-1422.

[16] W. Ghai and N. Singh, "Literature Review on Automatic Speech Recognition," in International Journal of Computer Applications, vol. 41, 2012, pp. 42-50.

[17] Y. X. Zou, W. Q. Zheng, W. Shi and H. Liu, "Improved Voice Activity Detection Based on Support Vector Machine With High Separable Speech Feature Vectors," in Proceedings of 19th International Conference on Digital Signal Processing, Hong Kong, China, 2014.

[18] F. Sheng, G. Zhang and Z. Song, "Comparison of Different Implementations of MFCC," in the Journal of Computer Science and Technology, vol. 16, Sept, 2001, pp. 582-589.

[19] W. Li and H. Bourlard, "Sub-Band Based Log-Energy and its Dynamic Range Stretching for Robust In-Car Speech Recognition," in Proceedings of the 13th Annual Conference of the International Speech Communication Association, Portland, OR, USA, September, 2012.

[20] E. M. Mugler, J. L. Patton, R. D. Flint, Z. A. Wright, S. U. Schuele, J. Rosenow, J. J. Shih, D. J. Krusienski and M. W. Slutzky, "Direct Classification of All American English Phonemes Using Signals from Functional Speech Motor Cortex," in Journal of Neural Engineering, vol. 11, June, 2014.

[21] Y. Su, F. Jelinek and S. Khudanpur, "Large-Scale Random Forest Language Models for Speech Recognition," in Proceedings of Interspeech, Antwerp, Belgium, 2007.

[22] J. Sohn and N. Kim, "A Statisitcal Model-based Voice Activity Detection," IEEE Signal Processing Letters, vol. 6, January, 1999, pp. 1-3.