IMPROVED DNN-BASED SEGMENTATION FOR MULTI-GENRE BROADCAST AUDIO

L. Wang, C. Zhang, P.C. Woodland, M.J.F. Gales, P. Karanasou, P. Lanchantin, X.Liu, Y. Qian

Cambridge University Engineering Dept, Trumpington St., Cambridge, CB2 1PZ U.K.

{lw519,cz277,pcw,mjfg,pk407,pkl27,xl207,yq236}@eng.cam.ac.uk

ABSTRACT

Automatic segmentation is a crucial initial processing step for processing multi-genre broadcast (MGB) audio. It is very challenging since the data exhibits a wide range of both speech types and background conditions with many types of non-speech audio. This paper describes a segmentation system for multi-genre broadcast audio with deep neural network (DNN) based speech/non-speech detection. A further stage of change-point detection and clustering is used to obtain homogeneous segments. Suitable DNN inputs, context window sizes and architectures are studied with a series of experiments using a large corpus of MGB television audio. For MGB transcription, the improved segmenter yields roughly half the increase in word error rate, over manual segmentation, compared to the baseline DNN segmenter supplied for the 2015 ASRU MGB challenge.

***Index Terms*—** audio segmentation, deep neural network, multi-genre broadcast data

1. INTRODUCTION

There has long been interest in processing streams of broadcast news data for automatic transcription and other applications. While broadcast news data is relatively predictable in terms of the data content and range, processing more general broadcast audio is much more challenging. Multi-genre broadcast (MGB) data, such as the data used in the 2015 MGB challenge task [1], can cover a wide range of genres, including news, comedy, drama, documentary, quiz shows, sports shows, etc. It typically includes audio data in very diverse environments, with speech in various speaking styles or over music, and multiple types of background noise or sound effects. Automatic segmentation is normally the first step used in processing broadcast audio, and for MGB data is a very challenging task.

Automatic audio segmentation aims to partition the speech regions in the audio data into homogeneous segments, with ideally one speaker and one background audio condition in each segment, so normalisation and adaptation based on segment clusters will be effective for following stages of transcription, diarisation etc. Two stages are usually included in the segmentation process: speech/non-speech detection to identify the speech regions, often referred to as voice activity detection (VAD), or speech activity detection (SAD), and speaker segmentation/clustering to obtain homogeneous segments.

Speech/non-speech detection is a fundamental task in almost all areas of speech technology, and it can have a significant impact on performance of speech recognition and speaker diarisation. Traditionally, low dimensional features such as energy, zero-crossing rate, periodicity measures and formant information[2, 3], were used with a threshold-based decision to detect the presence of speech. However, in challenging situations where non-speech segments might include various sorts of music and ambient noise, such as broadcast data, the general approach is to do frame-wise classification using more sophisticated features and models pre-trained using speech and non-speech data. Maximum likelihood classification with Gaussian mixture models (GMMs) is a popular model-based approach in these domains [4, 5, 6]. One speech model and one non-speech model were used in [7], while in [8, 9, 10] multiple speech models were used for the possible bandwidth and gender combinations. For non-speech models, noise and music are often explicitly modelled with labelled training data in many systems. In [11, 12, 13], there are classes for speech, music, noise, speech in music, and speech in noise, while in [8, 10], wide-band speech, narrow-band speech, music, and speech+music models are used. Classification models other than GMMs have also been investigated in speech region detection, such as support vector machine (SVM) [14, 15], conditional random fields (CRFs) [16], and neural networks [17]. Recently, deep learning approaches have become widely used in speech recognition [18, 19], and also show superiority over other classification models in speech/non-speech detection. In [20], a deep belief network (DBN) was used with multiple acoustic features, while in [21, 22], DNNs were explored for frame classification. Multiple DNNs and jointly trained DNNs were proposed in [23, 24] to tackle the issue of various noise conditions. Recurrent neural networks (RNNs) were also investigated in [25, 26].

Since DNN models have been proven to have high accuracy in classifying speech frames by learning complex patterns with deep structures, here we explore the use of DNN models in speech/non-speech detection for MGB data. The most suitable DNN inputs, context window sizes and architectures are extensively studied with a series of experiments. The use of a speaker segmentation/clustering stage, including speaker change-point detection (CPD) and bottom-up iterative agglomerative clustering (IAC) is examined and applied to the initial segments to ensure segment homogeneity.

The work in this paper is done in the context of developing a segmentation system for use in all four evaluation tasks (standard and longitudinal speech-to-text transcription [27], alignment [28] and longitudinal speaker diarisation and linking [29]) in the 2015 MGB challenge [1], which was an official challenge of ASRU 2015. The data used for the MGB challenge was supplied by the British Broadcasting Corporation (BBC) and consists of audio from BBC television programmes. The data covers a full range of genres and there are various types of non-speech included in it, such as music, applause, laughter, and various types of noise, which together pose significant challenges for automatic segmentation. Another key issue in dealing with the MGB challenge data is that only lightly su-

This work is in part supported by EPSRC Programme Grant EP/I031022/1 (Natural Speech Technology). Chao Zhang is also supported by Cambridge International Scholarship from the Cambridge Commonwealth, European & International Trust. Supporting data for this paper is available at https://www.repository.cam.ac.uk/handle/1810/253408.

pervised acoustic training data was available which is based on automatic alignments of subtitles (closed captions) for the BBC programmes. Hence care needs to be taken in order to obtain suitable training data for segmentation systems.

The rest of the paper is organised as follows. Section 2 describes the DNN-based speech/non-speech detection stage, and a series of experiments related to the design of DNN structures are presented in detail. The speaker segmentation/clustering stage is described in Section 3, and the segmenters trained with various different training sets are presented in Section 4. Finally, the segmentation performance and associated impact on transcription error rates of various segmenters for the MGB challenge standard transcription task are given in Section 5, and conclusions are drawn in Section 6.

2. DNN-BASED SPEECH/NON-SPEECH DETECTION

2.1. The DNN-HMM hybrid approach

A DNN is a multi-layer perceptron (MLP) with a number of hidden layers, and its input is usually formed from a stacked set of adjacent frames of the acoustic feature vector. The input to each unit in a hidden layer is the weighted sum of outputs from the previous layer, and each unit transforms the input with a hidden activation function such as the sigmoid function. Finally, in the output layer, the inputs to each unit are normalised to be the posterior probability of its associated class, typically using the softmax function. To interface a DNN with hidden Markov models (HMMs), the posterior probabilities are converted to the log-likelihood of an HMM state [19, 18].

We use this DNN-HMM hybrid approach for speech/non-speech detection. DNNs are trained with two softmax units in the output layer corresponding to speech and non-speech. The posterior probabilities are estimated by DNNs and converted to log-likelihood in the normal way, and frame-wise decisions are made in a Viterbi decoding framework with speech/non-speech HMMs with transition matrices that ensure a 2-frame minimum duration.

Layer-wise discriminative pre-training and fine-tuning [19] are employed in DNN training based on the frame-based cross-entropy (CE) criterion, and a random selection of 10% of the training data is held-out as validation data. Parameter updates are averaged over a mini-batch with 800 frames and smoothed by adding a "momentum" term of 0.5 times previous updates. The initial learning rate and minimum epoch number are set to 2.0×10^{-3} and 12. The following DNN segmentation and recognition experiments were performed using HTK 3.5 [30].

All DNNs in experiments in Section 2 use as input 40-dimensional filterbank features (FBK), which are increasingly popular in speech recognition and also used in the acoustic modelling stage of our transcription system for the MGB challenge [27], and the feature vector of current frame is extended with its preceding and succeeding frames. There are 6 hidden layers, and 200 sigmoid units in each hidden layer for the DNNs in Section 2.3. Experiments on the effects of differently sized input context windows and the use of various numbers of layers and hidden units are explored in Section 2.4.

2.2. Training data

The MGB Challenge provided participants with audio from seven weeks of BBC television programmes with a raw total of 1,600 hours of audio as the only acoustic training dataset. A lightly supervised decoding process [31] was applied to the audio to extract time boundaries for utterances, and details of data preparation for the challenge can be found in [1, 28]. The speech recogniser outputs for each recognised segment was compared to the aligned BBC subtitle transcripts, and an error rate computed between them was obtained at either the word level (a word matched error rate or WMER) or the phone level (phone matched error rate or PMER). The maximum WMER/PMER, along with an average word duration (AWD) threshold, was used to select data segments for training to ensure that the word/phone supervision information used for each speech segment selected is reasonably accurate.

The initial segmenters were trained using a 100h subset of the originally distributed training dataset. The segments chosen were randomly selected from those which had an AWD of less than 0.7s and WMER of less than 25% from lightly supervised alignment. The selected speech segments were aligned at the phone level and audio data of models that corresponded to speech was used as 'speech' data (62h) to train the initial DNN segmenters, and only intra-segment non-speech portions (silence and short pause) were used as 'non-speech' data (38h).

2.3. DNN input context window size

Previous research shows that longer context windows are beneficial for frame classification, and inputs of DNNs for speech recognition generally use a context window of several frames. For example, 11 frames were used in [18]. Thus, we have investigated performance of different sizes of input context windows, and the frame-level classification accuracy has been chosen as the evaluation measure in Table 1.

Input context window size	%Training accuracy	%Validation accuracy
9 frames	91.34	90.98
15 frames	92.51	92.08
23 frames	93.29	92.75
31 frames	93.44	92.87
39 frames	93.61	92.98
47 frames	93.70	93.07
55 frames	**93.78**	**93.09**
63 frames	93.80	93.08

Table 1. *Classification accuracies of DNNs with different sizes of input context windows.*

It can be seen that for the speech/non-speech classification task it is beneficial to use context windows that are considerably longer than those normally used for speech recognition with the best tested accuracy from a window covering more than 0.5s of audio. All subsequent experiments used a context window of 55 frames.

2.4. DNN structures

A series of experiments was done with different DNN structures, such as the number of hidden layers, and the number of sigmoid units in each hidden layer. Furthermore, we have also tried to increase the number of hidden units in the first hidden layer to make it more expressive for the very long input context windows. The classification performance of DNNs with all these configurations was given in Table 2.

From these experiments with different DNN architectures, we obtained a final DNN configuration: 40-dimensional filterbank features, 55 frames of input feature context, 6 hidden layers, 1,000 hidden units in the first hidden layer, 200 hidden units in other hidden layers, and a final output layer of two units to represent speech and non-speech. The DNN segmenter trained on the 100h training set as

DNN structure	%Training accuracy	%Validation accuracy
$2200 \times 200^2 \times 2$	93.64	92.99
$2200 \times 200^3 \times 2$	93.70	93.06
$2200 \times 200^4 \times 2$	93.73	93.07
$2200 \times 200^5 \times 2$	93.73	93.07
$2200 \times 200^6 \times 2$	93.78	93.09
$2200 \times 200^7 \times 2$	93.81	93.09
$2200 \times 200^8 \times 2$	93.82	93.09
$2200 \times 100^6 \times 2$	93.40	92.90
$2200 \times 400^6 \times 2$	94.13	93.05
$2200 \times 400 \times 200^5 \times 2$	94.05	93.09
$\mathbf{2200 \times 1000 \times 200^5 \times 2}$	**94.40**	**93.11**

Table 2. *Classification accuracies of DNNs with different structures. M^n means n hidden layers with M units in each layer.*

described in Section 2.2 with the above configuration, was denoted as DNN-v1.

3. SPEAKER SEGMENTATION AND CLUSTERING

After DNN-based speech/non-speech classification, the speech regions are labelled at the frame level. For effective normalisation and adaptation for speech recognition or other speaker-related processing, the detected speech regions should be partitioned into homogeneous segments through a speaker segmentation and clustering stage, with ideally one single speaker and one single background audio condition in each segment. The overall segmentation process is illustrated in Figure 1. The CPD and IAC stages draw on techniques from the Cambridge March 2005 system [10].

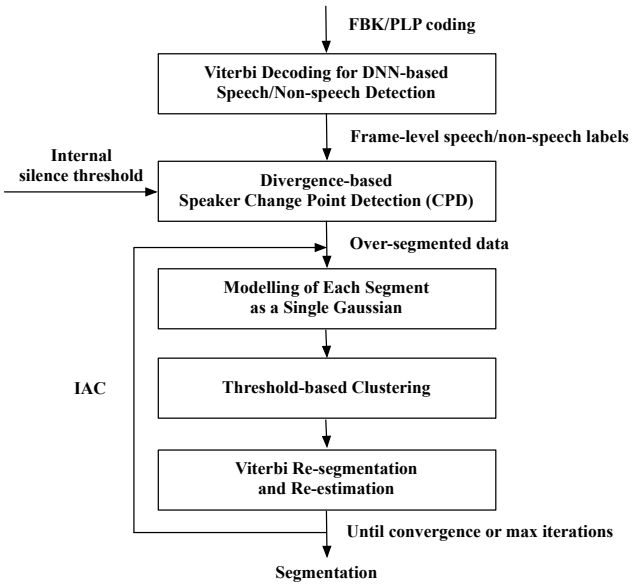

Fig. 1. Illustration of the overall segmentation process.

The CPD step finds potential changes in audio characteristics within each segment using the symmetric divergence distance metric (KL2) between two adjacent sliding windows of two seconds length. A full covariance Gaussian is used for each window, and a left-to-right search of local maxima is launched, removing the smaller of any pairs of neighbouring peaks occurring within a specified minimum duration. It was shown that enforcing a minimum length constraint of one second on the resulting segments reduced the segment impurity. A distance threshold is then chosen to over-segment the data. An internal silence threshold is set to discard larger silence portions and generate new segments with speech between these silences. The effect of the internal silence threshold on missed and false alarmed speech rates as well as word error rates (WER) of transcription systems will be shown in Section 5.

A bottom-up IAC step is then applied to the over-segmented data. For each iteration, a single Gaussian model is built for each segment and a new segment is formed by the two neighbouring segments with the smallest likelihood loss after merging. This procedure is repeated until a threshold of likelihood loss on merging is reached. Then re-segmentation of the data is performed by the Gaussian models using Viterbi decoding. The whole process is repeated until the segmentation converges or a maximum number of iterations are reached.

Finally, a maximum of 20 frames of silence (ensuring no overlapping segments) are added to the start and end of each segment to facilitate subsequent processing.

4. THE SEGMENTATION SYSTEMS

4.1. Refinement of the training dataset

The initial DNN-v1 based segmenter including CPD and IAC (the internal silence threshold was set to 50 frames) used the procedures in the previous sections, and the training data for DNN-v1 as detailed in Section 2.2. This segmenter was again applied to the whole MGB training set, which then went through another round of improved lightly supervised alignment with better acoustic and language models as described in [28]. The refined alignment resulted in a much larger amount of training data with PMER equal to zero, which therefore, allows us to use only this as a training set for frame-based DNN segmenters while giving sufficient training data and good frame alignment for speech/non-speech. This choice of data led to a 209h set of audio when AWD was limited to between 0.165s and 0.66s and PMER of zero was used. This data was then used as the basis for training further DNN-based segmenters.

4.2. Choosing background speech data

With the refined training data selection and alignments, two other segmenters, DNN-v3 and DNN-v4, both the same configuration as DNN-v1, were trained. Just as for DNN-v1, the audio data aligned to speech HMM states is used as 'speech' data (173h). DNN-v3 uses only audio data aligned to silence and short pause states inside a segment as 'non-speech' data (37h), while DNN-v4 uses a large sample of inter-segment non-speech data as well (313h). For DNN-v4, the potential speech portions outside aligned segments was filtered using a previously trained DNN segmenter[1], since it is known that some of the data not included in the BBC subtitles will in fact include speech data. It was found that the performance of the system is much poorer if a significant portion of data labelled as non-speech is actually speech.

The frame-level classification rates of these DNN segmenters are listed in Table 3 for different training and validation sets.

[1]For quick turnaround the segmenter used for background filtering was trained on the development set, but later experiments with a filtering DNN trained on the training set led to very similar performance.

Segmenter	%Training accuracy	%Validation accuracy
DNN-v1	94.40	93.11
DNN-v3	96.89	96.05
DNN-v4	**98.27**	**97.84**

Table 3. *Frame accuracies of DNNs with different training data sets.*

4.3. Testing different front-end features

Finally it was decided to investigate the use of other signal processing analyses besides filterbank features. Hence using the same training data and DNN architecture as those of DNN-v4, DNNs were trained using either 13-dimensional Mel frequency cepstral coefficients (MFCC) and 13-dimensional perceptual linear predictive (PLP). The results in Table 4 show that all front-ends perform well but the original filter bank system has a slight advantage.

Front-ends	%Training accuracy	%Validation accuracy
FBK	**98.27**	**97.84**
MFCC	98.11	97.79
PLP	98.09	97.74

Table 4. *Classification accuracies of DNNs with different front-ends.*

5. SEGMENTATION PERFORMANCE

5.1. Transcription system description

The performance of different segmenters were evaluated on **dev.full**, the 28h standard transcription task development test set for the MGB challenge with 47 different programme episodes. The transcription system used a speaker-independent (SI) sequence-trained DNN-HMM trained on a 700h subset of the original MGB training data (with PMER less than 40%) [27]. This DNN takes as input 40-dimensional filterbank features and their deltas with a context window of 9 frames, and has 5 hidden layers with 1,000 sigmoid units in each hidden layer, and 9.5k softmax units in the output layer. The 4-gram language model used is also detailed in [27].

5.2. Baseline segmentation systems

Baseline segmentation performance on **dev.full** is listed in Table 5.

Baselines	#Segs	%MS	%FA	%WER
Manual	**30,691**	-	-	**26.7**
LIUM	12,931	6.9	11.2	35.5
Cam RT-04	13,506	4.1	7.3	33.7
MGB baseline	**13,851**	**3.6**	**3.7**	**30.7**

Table 5. *Performance of baseline segmentations on dev.full.*

This includes the segmentation obtained from manual transcripts ('Manual'), the segmentation generated by the LIUM speaker diarization toolkit v8.4.1 using supplied models [32] ('LIUM'), the segmentation from the Cambridge RT-04 segmenter [33] ('Cam RT-04'), and the supplied automatic segmentation provided for the MGB challenge [34] ('MGB baseline'). The LIUM segmentation uses 13-dimensional MFCC features and 8 GMM models (including silence, speech, jingles, and music) for speech detection. The Cambridge RT-04 segmenter is based on GMM models trained on US broadcast news data only. The MGB baseline segmentation trained on MGB data is DNN-based and uses filterbank features of 23 dimensions with a context window of 15 frames on both sides. There are 2 hidden layers of 1,000 nodes and an output layer consisting of 2 nodes for speech (116h) and non-speech (363h).

For each segmentation, the number of segments, missed speech (MS) and false-alarmed speech (FA) rates as well as the transcription WERs are listed in Table 5[2]. The very large increase in transcription WER over the manual segmentation for the LIUM and Cambridge RT-04 shows how difficult the segmentation task is for this data. The DNN-based MGB baseline segmenter performs much better but still increases the WER by 4% absolute over the manual segments.

5.3. DNN-based segmentations

The performance of DNN-based segmentations, optionally with CPD and IAC steps, are shown in Table 6. Different internal silence thresholds in the CPD step were also investigated.

System	Max Sil	CPD IAC	#Segs	%MS	%FA	%WER
Manual	-	-	**30,691**	-	-	**26.7**
DNN-v1	50	-	18,926	2.1	5.5	30.4
DNN-v1	50	✓	17,191	2.6	4.2	29.9
DNN-v3	50	✓	17,332	1.7	5.4	30.0
DNN-v4	50	✓	15,529	2.2	2.1	29.1
DNN-v4	40	✓	17,363	2.3	2.0	28.9
DNN-v4	**30**	✓	**20,084**	**2.5**	**1.9**	**28.8**
DNN-v4	20	✓	23,860	2.7	1.8	28.8

Table 6. *Performance of different segmentations on dev.full. 'Max Sil' means the internal silence threshold in frames.*

With the use of additional inter-segment non-speech data, DNN-v4 achieved a slightly higher MS rate, but a much lower FA rate, compared to DNN-v3. Moreover, by reducing the internal silence threshold, more segments were generated and lower WERs were obtained with minor differences in MS and FA rates. It can be seen that the best performance comes from the segmentation system based on DNN-v4 with 30 frames of internal silence in the CPD step, which gave relative reduction of 30.7% and 48.6% in MS and FA rates and also a 1.9% reduction in WER over the MGB baseline. The degradation of this segmentation system from the manual segmentation over the MGB baseline was reduced from 4% absolute to 2.1% absolute.

6. CONCLUSION

This paper describes DNN-based automatic segmentation for multi-genre broadcast audio. By carefully tuning the DNN architecture and the training data used a much improved system results. Furthermore to obtain suitable segments for speech transcription, improved performance is obtained by including an extra change point detection and clustering stage. The final segmentation system roughly halved the increase in WER from automatic segmentation relative to the baseline DNN system supplied for the MGB challenge. This segmenter was used in all tasks (transcription, alignment, diarisation) for the Cambridge systems for the ASRU MGB challenge and was a key component in their good performance.

[2]Note that missed and false-alarmed speech rates used here for segmentation performance evaluation are slightly different to the missed and false-alarmed speaker rates commonly used for diarisation system evaluation.

7. REFERENCES

[1] P. Bell, M.J.F. Gales, T. Hain, J. Kilgour, P. Lanchantin, X. Liu, A. McParland, S. Renals, O. Saz, M. Wester & P.C. Woodland, "The MGB challenge: Evaluating multi-genre broadcast media transcription", *Proc. ASRU*, Scottsdale, 2015.

[2] R. Chengalvarayan, "Robust energy normalization using speech/nonspeech discriminator for German connected digit recognition", *Proc. Eurospeech*, Budapest, 1999.

[3] J.D. Hoyt & H. Wechsler, "Detection of human speech in structured noise", *Proc. ICASSP*, Adelaide, 1994.

[4] S.E. Tranter & D.A. Reynolds, "An overview of automatic speaker diarization systems", *IEEE Trans. ASLP*, Vol. 15, No. 5, pp. 1557-1565, 2006.

[5] X. Anguera, S. Bozonnet, N. Evans, C. Fredouille, G. Friedland & O. Vinyals, "Speaker diarization: a review of recent research", *IEEE Trans. ASLP*, Vol. 20, No. 2, pp. 356-370, 2012.

[6] M.H. Moattar & M.M. Homayounpour, "A review on speaker diarization systems and approaches", *Speech Communication*, Vol. 54, No. 10, pp. 1065-1103, 2012.

[7] C. Wooters, J. Fung, B. Peskin & X. Anguera, "Toward robust speaker segmentation: The ICSI-SRI Fall 2004 diarization system", *Proc. Fall Rich Transcription Workshop (RT-04)*, 2004.

[8] T. Hain, S.E. Johnson, A. Tuerk, P.C. Woodland & S.J. Young, "Segment generation and clustering in the HTK broadcast news transcription system", *Proc. DARPA Broadcast News Transcription and Understanding Workshop*, pp. 133-137, 1998.

[9] P. Nguyen, L. Rigazio, Y. Moh &J.C. Junqua, "Rich transcription 2002 site report. Panasonic speech technology laboratory (PSTL)", *Proc. Rich Transcription Workshop (RT-02)*, 2002.

[10] R. Sinha, S.E. Tranter, M.J.F. Gales & P.C. Woodland, "The Cambridge University March 2005 speaker diarization system", *Proc. Eurospeech*, Lisbon, 2005.

[11] J.-L. Gauvain, L. Lamel & G. Adda, "Partitioning and transcription of broadcast news data", *Proc. ICSLP*, Sydney, 1998.

[12] D.A. Reynolds & P. Torres-Carrasquillo, "The MIT Lincoln Laboratory RT-04F diarization systems: Applications to broadcast audio and telephone conversations", *Proc. Fall 2004 Rich Transcription Workshop (RT-04)*, 2004.

[13] X. Zhu, C. Barras, S. Meignier & J.-L. Gauvain, "Combining speaker identification and BIC for speaker diarization", *Proc. Eurospeech*, Lisbon, 2005.

[14] N. Mesgarani, M. Slaney & S.A. Shamma, "Discrimination of speech from nonspeech based on multiscale spectro-temporal modulations", *IEEE Trans. ASLP*, Vol. 14, No. 3, pp. 920-930, 2006.

[15] J.W. Shin, J.H. Chang & N.S. Kim, "Voice activity detection based on statistical models and machine leaning approaches", *Computer Speech and Language*, Vol. 24, No. 3, pp. 515-530, 2010.

[16] A. Saito, Y. Nankaku, A. Lee & K. Tokuda, "Voice activity detection based on conditional random fields using multiple features", *Proc. Interspeech*, Makuhari, 2010.

[17] A. Bugatti, A. Flammini & P. Migliorati, "Audio classification in speech and music: a comparison between a statistical and a neural approach", *EURASIP Journal on Applied Signal Processing*, No. 1, pp. 372-378, 2002.

[18] G.E. Dahl, D. Yu & A. Acero, "Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition", *IEEE Trans. ASLP*, Vol. 20, No. 1, pp. 30-42, 2012.

[19] G. Hinton, L. Deng, D. Yu, G.E. Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T.N. Sainath & B. Kingsbury, "Deep neural networks for acoustic modeling in speech recognition", *IEEE Signal Processing Magazine*, Vol. 29, No. 6, pp. 82-97, 2012.

[20] X.-L. Zhang & J. Wu, "Deep belief networks based voice activity detection", *IEEE Trans ASLP*, Vol. 21, No. 4, pp. 697-710, 2013.

[21] N. Ryant, M. Liberman & J. Yuan, "Speech activity detection on youtube using deep neural networks", *Proc. Interspeech*, Lyon, 2013.

[22] X.-L. Zhang & J. Wu, "Boosted deep neural networks and multi-resolution cochleagram features for voice activity detection", *Proc. Interspeech*, Singapore, 2014.

[23] I. Hwang, J. Sim, S.-H. Kim, K.-S. Song & J.-H. Chang, "A statistical model-based voice activity detection using multiple DNNs and noise awareness", *Proc. Interspeech*, Dresden, 2015.

[24] Q. Wang, J. Du, X. Bao, Z.-R. Wang, L.-R. Dai & C.-H. Lee, "A universal VAD based on jointly trained deep neural networks", *Proc. Interspeech*, Dresden, 2015.

[25] T. Hughes, & K. Mierle, "Recurrent neural networks for voice activity detection", *Proc. ICASSP*, Vancouver, 2013.

[26] F. Eyben, F. Weninger, S. Squartini & B. Schuller, "Real-life voice activity detection with LSTM recurrent neural networks and an application to Hollywood movies", *Proc. ICASSP*, Vancouver, 2013.

[27] P.C. Woodland, X. Liu, Y. Qian, C. Zhang, M.J.F. Gales, P. Karanasou, P. Lanchantin & L. Wang, "Cambridge University transcription systems for the Multi-Genre Broadcast challenge", *Proc. ASRU*, Scottsdale, 2015.

[28] P. Lanchantin, M.J.F. Gales, P. Karanasou, X. Liu, Y. Qian, L. Wang, P.C. Woodland & C. Zhang, "The development of the Cambridge University alignment systems for the Multi-Genre Broadcast challenge", *Proc. ASRU*, Scottsdale, 2015.

[29] P. Karanasou, M.J.F. Gales, P. Lanchantin, X. Liu, Y. Qian, L. Wang, P.C. Woodland & C. Zhang, "Speaker diarisation and longitudinal linking in multi-genre broadcast data", *Proc. ASRU*, Scottsdale, 2015.

[30] C. Zhang & P.C. Woodland, "A general artificial neural network extension for HTK", *Proc. Interspeech*, Dresden, 2015.

[31] H.Y. Chan & P.C. Woodland, "Improving broadcast news transcription by lightly supervised discriminative training", *Proc. ICASSP*, Montreal, 2004.

[32] M. Rouvier, G. Dupuy, P. Gay, E. Khoury, T. Merlin & S. Meignier, "An open-source state-of-the-art toolbox for broadcast news diarization", *Proc. Interspeech*, Lyon, 2013.

[33] M.J.F. Gales, D.Y. Kim, P.C. Woodland, H.Y. Chan, D. Mrva, R. Sinha & S.E. Tranter. "Progress in the CU-HTK broadcast news transcription system", *IEEE Trans. ASLP*, Vol. 14, No. 5, pp. 1513-1525, 2006.

[34] R. Milner, O. Saz, S. Deena, M. Doulaty, R.W.M. Ng & T. Hain, "The 2015 Sheffield system for longitudinal diarisation of broadcast media", *Proc. ASRU*, Scottsdale, 2015.