Complete-Linkage Clustering for Voice Activity Detection in Audio and Visual Speech

Houman Ghaemmaghami, David Dean, Shahram Kalantari, Sridha Sridharan, Clinton Fookes

Speech and Audio Research Laboratory, Science and Engineering Faculty,
Queensland University of Technology, Brisbane, Australia

houman.ghaemmaghami@qut.edu.au, ddean@ieee.org, s1.kalantari@qut.edu.au,
s.sridharan@qut.edu.au, c.fookes@qut.edu.au

Abstract

We propose a novel technique for conducting robust voice activity detection (VAD) in high-noise recordings. We use Gaussian mixture modeling (GMM) to train two generic models; speech and non-speech. We then score smaller segments of a given (unseen) recording against each of these GMMs to obtain two respective likelihood scores for each segment. These scores are used to compute a dissimilarity measure between pairs of segments and to carry out complete-linkage clustering of the segments into speech and non-speech clusters. We compare the accuracy of our method against state-of-the-art and standardised VAD techniques to demonstrate an absolute improvement of 15% in half-total error rate (HTER) over the best performing baseline system and across the QUT-NOISE-TIMIT database. We then apply our approach to the Audio-Visual Database of American English (AVDBAE) to demonstrate the performance of our algorithm in using visual, audio-visual or a proposed fusion of these features.

Index Terms: Voice activity detection, high noise, Gaussian mixture modeling, complete-linkage clustering

1. Introduction

Voice activity detection (VAD) is the process of identifying periods of active speech in a given recording and is a necessary front-end module to nearly all speech processing applications. It is commonly used to discard the non-speech periods of a recording prior to conducting speech recognition [1], speaker diarization [2], speaker recognition [3] or speech coding [4]. It can be used for estimating the noise spectrum in speech enhancement applications [5], or to simply reduce the time and effort required to listen to large sets of spoken recordings in human listening applications.

The proposed VAD techniques in the literature typically consist of a feature extraction stage followed by speech/non-speech classification. Some of the common features used for VAD include energy, zero-crossing rate, cepstral coefficients, autocorrelation features [6] and spectral divergence [7]. Most VAD techniques use some combination of these. The typical classification methods employed in recent VAD algorithms include simple techniques, such as heuristic-based approaches using tunable thresholds [8], or more complex classifiers such as deep neural networks (DNN) [9], hidden Markov modeling (HMM) with Gaussian mixture models (GMM) [2], support vector machines (SVM) [10] and Gaussian likelihood ratio testing (LRT) [11]. There are a large variety of techniques in the literature that display inadequate performance in high-noise recording scenarios and a clear lack of an all-round, noise-robust VAD approach that can reliably be applied to multiple recording domains and high-noise recordings - with signal-to-noise ratios (SNR) of < 5 dB. At the same time, the need for noise-robust and efficient audio processing is increasing, particularly in the fields of speech and speaker recognition [12, 13]. The development of such a VAD algorithm has been the motivation for this work.

We propose a novel VAD approach for detecting speech in noisy recordings and across a wide range of noise scenarios. We first train two GMMs for representing the distribution of speech and non-speech features in the training set, respectively. For VAD on a previously unseen recording, we use our trained GMMs to compute speech and non-speech likelihood scores for smaller segments of the given recording. We then draw from our work on speaker linking and clustering [14], to propose a complete-linkage clustering classifier for conducting VAD using the likelihood scores. We propose a pairwise dissimilarity measure to compare segments and to cluster them into classes of speech or non-speech using complete-linkage clustering. We evaluate our algorithm on the QUT-NOISE-TIMIT database [15] and against the performance of five baseline VAD systems: ITU-T G.729 Annex B [4], advanced front-end (AFE) ETSI [1] long term spectral divergence (LTSD) [7], Sohns likelihood ratio test (LRT) VAD [11] and a GMM based learning approach using mel-frequency cepstral coefficient features (GMM-MFCC) [15]. We demonstrate that our proposed system outperforms the best baseline VAD by up to 15% in absolute half-total error rate (HTER) [6]. We then employ the Audio-Visual Database of American English (AVDBAE) [16] to demonstrate our VAD technique using visual or audio-visual features, without the need for system tuning or thresholding. Finally, we propose a score fusion approach for carrying out audio-visual VAD (AV-VAD) with our system and show that we can improve VAD performance using this fusion scheme.

2. Relation to prior work

One of the main shortcomings of prior work on voice activity detection (VAD) is the lack of adequate performance evaluations that are conducted on a wide range of real noise recordings (SNR < 5 dB) [15]. In addition, most VAD techniques require some form of threshold tuning or calibration to achieve their reported performance [8, 7]. Our work in this paper demonstrates a highly noise-robust VAD algorithm that can be applied to multiple audio domains without any necessary tuning or thresholding. We evaluate our technique over a wide variety of real noisy recording scenarios (600 hours at 10 noise locations) and at SNR levels as low as -10 dB, to demonstrate a 15% absolute

improvement in error rates over our best performing baseline system. In addition, many VAD studies have often focused on engineering specific VAD features for this task [7, 10, 6]. Our approach does not require a specific feature recipe. We demonstrate this by using audio, visual or audio-visual features for VAD. In addition, we build upon our previous work on visual VAD [17], to present a novel score fusion scheme for taking advantage of audio and visual features for greater accuracy in audio-visual VAD (AV-VAD) [18].

3. VAD algorithm

In our proposed voice activity detection (VAD) approach we use two previously trained Gaussian mixture models (GMM) (detailed in Section 4), one for speech and another for non-speech, to compute two respective likelihood scores for segments of a given recording. We then use the segment likelihood scores to calculate a pairwise dissimilarity measure between segments. This dissimilarity measure is used to cluster the segments into speech and non-speech clusters, achieving a final VAD decision. It is important to note that the data used for training the speech and non-speech GMMs does not overlap with the evaluation data in any of the evaluations presented in this paper. The training process and evaluations are detailed in Section 4.

3.1. Likelihood scoring

We begin VAD of a recording using our previously trained speech and non-speech GMMs. We first extract common mel-frequency cepstral coefficient (MFCC) features from the audio and split the set of feature vectors into smaller segments $\boldsymbol{X}$, where $\boldsymbol{X} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_K\}$ is a set of multi-dimensional feature vectors. The two GMMs can then be used to obtain the likelihood of a segment $\boldsymbol{X}$ being speech or non-speech.

GMMs can be used to model an arbitrarily-shaped continuous density by using a sufficient number of Gaussian densities and through adjusting the means, covariances and weights for each of these Gaussian components [19]. We use an equal number of component densities (C) for training both the speech and non-speech GMMs. The log-likelihood (LL) of a segment $\boldsymbol{X}$ given a GMM (θ) can then be computed as,

$$\log P(\boldsymbol{X}|\theta) = \sum_{k=1}^{K} \log p(\boldsymbol{x}_k|\theta) = \sum_{k=1}^{K} \log \sum_{c=1}^{C} \omega_c g(\boldsymbol{x}_k|\boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c)\,, \tag{1}$$

where g represents a Gaussian distribution function, ω_C represents the mixture weights, $\boldsymbol{\mu}_C$ the mean vectors and $\boldsymbol{\Sigma}_C$ the covariance matrices of the C mixture components.

For every segment $\boldsymbol{X}$ we obtain two log-likelihood scores: the log-likelihood score of $\boldsymbol{X}$ against the speech GMM referred to as $LL_s(\boldsymbol{X})$, and the log-likelihood of $\boldsymbol{X}$ given the non-speech GMM referred to as $LL_n(\boldsymbol{X})$. This provides the clustering stage (Section 3.2) of our algorithm with a set of two log-likelihood scores for each segment $\boldsymbol{X}$.

3.2. Complete-linkage clustering

We use complete-linkage clustering to cluster segments based on their speech and non-speech log-likelihood scores extracted in Section 3.1. Complete-linkage clustering has previously been used for speaker clustering in the context of speaker linking and diarization [20]. Linkage clustering is a form of agglomerative clustering that employs a *linkage* rule to update pairwise segment scores after a merge, thus allowing for efficient clustering [21]. To do this, we need to compute a pairwise dissimilarity measure between all pairs of segments. We define a simple dissimilarity measure between two segments, $\boldsymbol{X}_i$ and $\boldsymbol{X}_j$ as,

$$d(i,j) = \frac{1}{|LLR(\boldsymbol{X}_i) + LLR(\boldsymbol{X}_j)|}, \tag{2}$$

where $LLR(\boldsymbol{X}_i)$ is the log-likelihood ratio for segment $\boldsymbol{X}_i$ and the log-likelihood ratio for any segment $\boldsymbol{X}$ is computed as,

$$LLR(\boldsymbol{X}) = LL_s(\boldsymbol{X}) - LL_n(\boldsymbol{X}). \tag{3}$$

From (2) and (3), $d(i,j)$ will be lower for similar segments (both speech or both non-speech), while it would be higher for dissimilar segments (one speech and one non-speech).

After obtaining the pairwise dissimilarity scores d, we begin the complete-linkage clustering process by first merging the most similar pair of segments to form a starting node. The pairwise measure d between this new node and each of the remaining segments is then updated to reflect the most dissimilar score between any of their elements. For example, if we merge two segments $\boldsymbol{X}_i$ and $\boldsymbol{X}_j$ into $\boldsymbol{X}'_i = \{\boldsymbol{X}_i, \boldsymbol{X}_j\}$, the score between the newly formed cluster of segments $\boldsymbol{X}'_i$ and any other segment $\boldsymbol{X}_n$ will be updated to $d(i', n)$ where,

$$d(i', n) = \max(d(i,n), d(j,n)). \tag{4}$$

Complete-linkage clustering provides a cautious clustering rule that will consistently take into account the worst-case score scenario after every segment merge. We use complete-linkage clustering to cluster all segments, based on their pairwise dissimilarity scores d defined in (2), down to two final clusters $\boldsymbol{C}_1$ and $\boldsymbol{C}_2$. After achieving two dissimilar clusters, we then need to decide which is the speech cluster and which is the non-speech. We apply a simple relative comparison test using (3) which is the log-likelihood ratio of a segment being speech; if $LLR(\boldsymbol{C}_1) > LLR(\boldsymbol{C}_2)$, then $\boldsymbol{C}_1$ is more likely (than $\boldsymbol{C}_2$) to be the speech cluster, if not then $\boldsymbol{C}_2$ is more likely to be speech. This clustering and relative detection approach eliminates the need for any decision thresholding or parameter tuning when applying our proposed VAD approach across multiple audio domains. Of course, this implies that we assume the existence of active speech in every processed recording, which may be a sound assumption in some speech recognition and speaker verification tasks. However, in scenarios where having no active speech or no non-speech segments is a possibility, one can simply append a short dummy speech and non-speech segment to the beginning of processed recordings to ensure the existence of active speech and then ignore the VAD decisions for the portion of the recording containing the dummy segment. Incorporating the reference knowledge regarding the nature of the dummy segments can also be used to guide the clustering process. This is possible in the case of our proposed algorithm as it employs a relative comparison to separate the recording into speech and non-speech clusters, which eliminates the need for tuning and thus ensures its robustness across varied audio domains.

After obtaining VAD decisions that indicate speech and non-speech portions of a recording, we apply a decision hangover based on our previous work [6]. We add 300 ms of preceding speech and 500 ms of proceeding speech to every detected speech segment, while removing segments shorter than 250 ms if no speech event is present within their hangover period. We do this to smooth and extend speech events and remove spurious speech decisions.

4. Noise-robustness evaluations

We utilise the QUT-NOISE-TIMIT database for VAD evaluations [15]. This corpus contains noisy recordings over various signal-to-noise ratio (SNR) levels and noise scenarios. We compare the performance of our proposed system to the evaluation results of five baseline VAD techniques over QUT-NOISE-TIMIT. These include two standardized off-the-shelf VAD systems: ITU-T G.729 Annex B [4] and advanced front-end (AFE) ETSI [1]. Our implementation of two state-of-the-art baseline VAD techniques: long term spectral divergence (LTSD) [7] and Sohns likelihood ratio test (LRT) VAD [11]. As well as a GMM based learning approach using MFCC features (GMM-MFCC) that is proposed and detailed in the QUT-NOISE-TIMIT evaluation protocol [15]. It must be noted that the training of the speech and non-speech GMMs is only carried out once prior to the evaluations and that the data employed for training does not in any way overlap (with respect to recording content or noise scenarios) with the evaluation data. This is done to ensure an unbiased evaluation. The experimental results are provided in this section.

4.1. QUT-NOISE-TIMIT corpus and experiments

The QUT-NOISE-TIMIT dataset is designed for training and extensive performance evaluation of VAD algorithms across high noise recordings. This dataset was constructed by mixing clean speech from the TIMIT corpus [22] with an extensive range of real background noise recordings from the QUT-NOISE corpus [15]. This dataset contains a total of 600 hours of noisy spoken recordings, created across 24,000 files, with 200 files for one of six SNR levels, for each of 20 recording sessions in the QUT-NOISE corpus. Figure 1 demonstrates the structure of the QUT-NOISE-TIMIT corpus.

We follow the exact training and testing protocol recommended for the QUT-NOISE-TIMIT corpus [15]. This dataset provides two independent and equal-sized sets of noisy recordings to allow for cross training and evaluation: Groups A and B. From Figure 1, we will train speech and non-speech GMMs on Group A to evaluate on Group B recordings, and vice versa. This is also applicable to LTSD and Sohn VAD systems, which were similarly trained (or tuned) on one Group and tested on the other. We use 19 MFCC features including the zeroth order coefficient with deltas and feature warping [23], extracted every 32 ms using a Hamming window and a 10 ms window shift. In addition, we use 16 mixture components to train our speech and non-speech GMMs. For likelihood segment scoring we use 5 feature vectors per segment, equivalent to a 50 ms.

4.2. Evluation results

We employ commonly used performance metrics to evaluate our algorithm and baseline systems [15]. These are the false alarm rate (FAR), miss rate (MR) and half-total error rate (HTER) metrics. FAR is the total time that a system erroneously makes speech decisions when there is no speech, over the total length of non-speech regions in a recording. MR is the total time that a system erroneously misses speech decisions, over the total length of the true speech events in a recording. We express these metrics as percentages and compute the HTER as the equal-weighted average of FAR and MR percentages.

Figure 2 displays the performance of our proposed VAD algorithm against the four baseline systems at each of three noise levels: low noise (SNR = 10 and 15 dB), medium noise (SNR = 0 and 5 dB) and high noise (SNR = -10 and -5 dB). In Figure 2 the HTER metric is displayed as a bar graph for each system and noise level. This is the overall HTER metric calculated across the entire dataset. Each bar is shaded, with dark and light shaded portions indicating the contribution of MR and FAR metrics to the overall HTER, respectively.

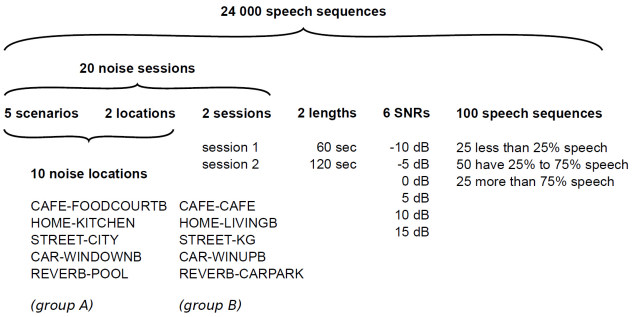

Figure 1: An overview of the structure of the QUT-NOISE-TIMIT corpus of noisy speech recordings [15].

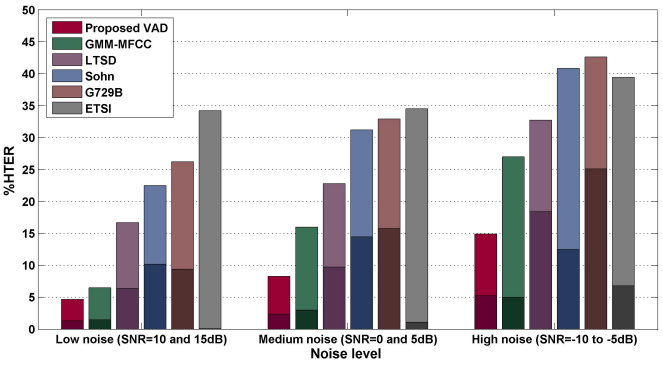

Figure 2: Evaluation of VAD systems across QUT-NOISE-TIMIT; the dark and light shaded portions of HTER bars represent the contribution of MR and FAR metrics, respectively.

Our proposed VAD outperforms the evaluated baseline systems. This is particularly noticeable in the high noise evaluations, where our algorithm outperforms the best baseline technique (GMM-MFCC) by an absolute value of 15% in HTER. In addition, unlike the state-of-the-art techniques such as LTSD, Sohn and GMM-MFCC VAD systems, our system does not employ any form of thresholding or tuning, nor does it depend on a specifically designed feature like the LTSD approach [7].

5. Audio-visual VAD

In the task of speaker recognition in multimedia datasets, it is often necessary to use multiple features in order to tie together the face and speech of a person speaking [24]. We have previously carried out work on visual VAD using a GMM based learning approach [17]. We now extend our work to audio-visual VAD (AV-VAD) and evaluate our proposed VAD approach using the Audio-Visual Database of American English (AVDBAE) [16]. This dataset contains relatively clean speech and as we have shown that our system is significantly more accurate than the baseline techniques in Section 4, we will hereon use our proposed algorithm to demonstrate VAD performance.

Table 1: VAD evaluations using different features across the AVDBAE corpus, where the audio-visual features are obtained by appending visual features to their respective audio features.

Feature type	HTER%	FAR%	MR%
Audio	**3.9**	7.1	0.7
Visual	**9.7**	17.9	1.5
Audio + Visual	**6.2**	6.9	5.5

In this section, we first conduct VAD using only audio features, only visual features or these features appended to form audio-visual features. We demonstrate that our proposed VAD performs best using only audio features. We then present a score fusion scheme for incorporating information from both audio and visual features for AV-VAD using our system to show improvements over the audio-only VAD accuracy.

5.1. AVDBAE corpus and experiments

We use the Audio-Visual Database of American English (AVDBAE), which contains 14 speakers; 10 female (F02-F011) and 4 male (M01-M04) with approximately equal data for each speaker [16]. We split this dataset into two non-overlapping sets for cross training and evaluation. To do this, we select all even numbered female and male speakers to form Group A = {F02, F04, F06, F08, F010, M02, M04}, with the remainder of the dataset forming Group B. As before, we train 16 mixture GMMs on Group A to test on Group B, and vice versa.

For visual features, we extract the mean-removed lip region-of-interest (ROI) for every frame of video at 29.97 fps [25]. We then apply a two-dimensional discrete cosine transform (DCT) to the mean-removed ROI, retaining the top 100 DCT coefficients according to the zigzag pattern, thus achieving a *static* visual feature vector. To extract dynamic speech information, we apply inter-frame linear discriminant analysis (LDA) to 7 consecutive frames (centered at the analysed frame) to obtain a 60 dimensional LDA feature vector. Finally, we apply feature warping to the extracted LDA features [23]. For audio analysis we use MFCC features as in Section (4.1), but extracted at the video frame rate of 29.97 fps with no overlap. This is to simplify the audio-visual feature fusion evaluations in this section. For likelihood scoring (Section 3.1), we use 2 vectors per segment to achieve a length close to 50 ms.

5.2. Evaluation results

We apply our VAD approach to audio, visual and audio-visual features. We obtain the audio-visual features by appending visual feature vectors to their respective audio feature vectors. From Table 1, our proposed VAD provides enough robustness to accommodate various features without tuning. It appears that appending the visual features to audio features slightly decreases FAR while raising MR. To further explore this, we propose a weighted log-likelihood score fusion approach to combine audio and visual segment likelihoods. Given a segment of audio features $\boldsymbol{X}_a$ and their respective visual features $\boldsymbol{X}_v$, we modify (3) for weighted log-likelihood score fusion,

$$\begin{aligned} LLR(\boldsymbol{X}_{av}) = &\big(\alpha LL_s(\boldsymbol{X}_a) + (1-\alpha) LL_s(\boldsymbol{X}_v)\big) \\ &-\big(\beta LL_n(\boldsymbol{X}_a) + (1-\beta) LL_n(\boldsymbol{X}_v)\big), \end{aligned} \quad (5)$$

where α and β are the speech and non-speech likelihood weighting factors, respectively, which range between 0.0 and 1.0 in steps of 0.1. From (5), α=β=1.0 would represent the audio-only experiments (Table 1) while setting both factors to 0.0 would represent the visual-only evaluations. We look at three fusion cases where we vary both, or at least one of the two weighting factors: ($\alpha = \beta = [0.0, 1.0]$), ($\alpha = 1.0, \beta = [0.0, 1.0]$) and ($\alpha = [0.0, 1.0], \beta = 1.0$). As the audio features had the best performance, when we are not varying a factor we set it to 1.0 to favour audio features. The results are shown in Figure 3. It can be seen that our weighted score fusion approach outperforms the audio-only results (Table 1) at ($\alpha = 0.9, \beta = 1.0$). This result suggests that the visual features may retain non-speech information which can be used to improve VAD, thus opening up additional opportunities for further investigation and improvement.

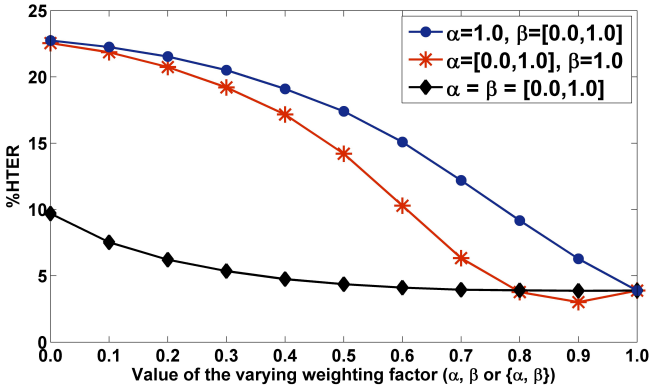

Figure 3: Weighted log-likelihood score fusion evaluations where α, β, or both factors ($\{\alpha, \beta\}$) are varied along the x-axis with 0.0 weighting factor values representing visual-only performance and values of 1.0 indicating audio-only performance.

6. Conclusion

We proposed a noise-robust VAD, which we evaluated against five baseline systems to demonstrate an absolute improvement of 15% in error rate, over the best baseline VAD, across QUT-NOISE-TIMIT dataset of noisy speech recordings. We showed that through employing complete-linkage clustering we can achieve threshold independence and outperform equivalent GMM-based learning techniques for conducting VAD in high-noise recording scenarios. We emphasise that no form of thresholding was used in our proposed algorithm and that the pre-trained speech and non-speech GMMs were trained on data that did not overlap with our evaluation set (not even with respect to noise type). We then used our proposed system to study its performance in carrying out audio-visual VAD over the AVDBAE corpus. To do this, we first evaluated our technique using audio-only, visual-only and concatenated audio and visual features to show that the best performance in this manner is achieved using audio-only features. We then proposed a novel score fusion approach for incorporating audio and visual feature information into our VAD scheme for AV-VAD and showed that we can outperform the best VAD performance, which was achieved using audio-only features, across the AVDBAE corpus.

7. Acknowledgements

This research was supported by an Australian Research Council (ARC) Linkage Grant (No: LP130100110).

8. References

[1] J.-Y. Li, B. Liu, R.-H. Wang, and L.-R. Dai, "A complexity reduction of ETSI advanced front-end for DSR," in *ICASSP 2004*, vol. 1, 17-21 2004, pp. I – 61–4 vol.1.

[2] C. Wooters and M. Huijbregts, "The ICSI RT07s speaker diarization system," in *Multimodal Technologies for Perception of Humans*. Springer Berlin / Heidelberg, 2008.

[3] D. A. Reynolds, "An overview of automatic speaker recognition technology," in *ICASSP 2002*, vol. 4, may 2002, pp. IV–4072 – IV–4075.

[4] A. Benyassine, E. Shlomot, S. H, D. Massaloux, C. Lambline, and J. Petit, "ITU-T Rec. G.729 Annex B: A silence compression scheme for G.729 optimized for V.70 digital simultaneous voice and data applications," ITU-T Recommendation G.729 Annex B, ITU, Tech. Rep., 1996.

[5] M. Marzinzik and B. Kollmeier, "Speech pause detection for noise spectrum estimation by tracking power envelope dynamics," *IEEE Transactions on Speech and Audio Processing*, vol. 10, no. 2, pp. 109–118, February 2002.

[6] H. Ghaemmaghami, B. B. B., R. Vogt, and S. Sridharan, "Noise robust voice activity detection using features extracted from the time-domain autocorrelation function," in *Interspeech 2010*, Makuhari Messe International Convention Complex, Makuhari, Japan, 2010. [Online]. Available: http://eprints.qut.edu.au/40656/

[7] J. Ramírez, J. Segura, C. Benítez, A. de la Torre, and A. Rubio, "Efficient voice activity detection algorithms using long-term speech information," *Speech Communication*, vol. 42, pp. 271–287, 2004.

[8] H. Ghaemmaghami, D. B. Dean, S. Sridharan, and I. Mccowan, "Noise robust voice activity detection using normal probability testing and time-domain histogram analysis," in *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*. Sheraton Dallas, Dallas, Texas: IEEE, 2010, pp. 4470–4473. [Online]. Available: http://eprints.qut.edu.au/40252/

[9] X.-L. Zhang, "Unsupervised domain adaptation for deep neural network based voice activity detection," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, May 2014, pp. 6864–6868.

[10] J. Ramirez, P. Yelamos, J. Gorriz, and J. Segura, "SVM-based speech endpoint detection using contextual speech features," *Electronics Letters*, vol. 42, no. 7, pp. 426 – 428, 30 2006.

[11] J. Sohn, N. Kim, and W. Sung, "A statistical model-based voice activity detection," *IEEE Signal Processing Letters*, vol. 6, no. 1, pp. 1–3, January 1999.

[12] V. Mitra, H. Franco, M. Graciarena, and A. Mandal, "Normalized amplitude modulation features for large vocabulary noise-robust speech recognition," in *ICASSP 2012*, March 2012, pp. 4117–4120.

[13] Y. Lei, L. Burget, and N. Scheffer, "A noise robust i-vector extractor using vector Taylor series for speaker recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, May 2013, pp. 6788–6791.

[14] H. Ghaemmaghami, D. Dean, and S. Sridharan, "Speaker linking using complete-linkage clustering," in *SST2012*, 2012.

[15] D. Dean, S. Sridharan, R. Vogt, and M. Mason, "The QUT-NOISE-TIMIT corpus for the evaluation of voice activity detection algorithms," in *Interspeech 2010*, Makuhari, Japan, September 2010. [Online]. Available: http://eprints.qut.edu.au/38144/

[16] S. Richie, C. Warburton, and M. Carter, "Audiovisual Database of spoken American English," in *Linguistic Data Consortium*, 2009.

[17] R. Navarathna, D. B. Dean, S. Sridharan, C. B. Fookes, and P. J. Lucey, "Visual voice activity detection using frontal versus profile views," in *The International Conference on Digital Image Computing : Techniques and Applications (DICTA2011)*, Sheraton Noosa Resort & Spa, Noosa, QLD, October 2011. [Online]. Available: http://eprints.qut.edu.au/46513/

[18] T. Yoshida and K. Nakadai, "Audio-visual voice activity detection based on an utterance state transition model," *Advanced Robotics*, vol. 26, no. 10, pp. 1183–1201, 2012.

[19] C. M. Bishop, *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Secaucus, NJ, USA: Springer-Verlag New York, Inc., 2006.

[20] H. Ghaemmaghami, D. Dean, R. Vogt, and S. Sridharan, "Speaker attribution of multiple telephone conversations using a complete-linkage clustering approach," in *IEEE ICASSP2012*, march 2012, pp. 4185 –4188.

[21] A. Jain, A. Topchy, M. Law, and J. Buhmann, "Landscape of clustering algorithms," in *Proceedings of ICPR2004*, vol. 1, 2004, pp. 260 – 263 Vol.1.

[22] W. Fisher, G. Doddington, and K. Goudie-Marshall, "The DARPA speech recognition research database: specifications and status," in *Proc. DARPA Workshop on Speech Recognition*, 1986, pp. 93–99.

[23] J. Pelecanos and S. Sridharan, "Feature warping for robust speaker verification," in *Odyssey2001*, June 18-22 2001, pp. 213–218.

[24] F. Battisti, M. Carli, M. Leo, and A. Neri, "Probabilistic person identification in TV news programs using image web database," in *IS&T/SPIE Electronic Imaging*. International Society for Optics and Photonics, 2014, pp. 90 190D–90 190D.

[25] S. Kalantari, R. Navarathna, D. B. Dean, and S. Sridharan, "Visual front-end wars : Viola-jones face detector vs fourier lucas-kanade," in *International Conference on Auditory Visual Speech Processing 2013*, B. Denis and B. Jonas, Eds., Ternélia resort Le Pré du Lac, Annecy, France, 2013. [Online]. Available: http://eprints.qut.edu.au/62749/