



Incorporating visual information for spoken term detection

Shahram Kalantari, David Dean, Sridha Sridharan

Speech, Audio, Image, and Video Technologies, Science and Engineering Faculty
 Queensland University of Technology, 2 George St, Brisbane, Australia 4000

s1.kalantari@qut.edu.au, ddean@ieee.org, s.sridharan@qut.edu.au

Abstract

Spoken term detection (STD) is the task of looking up a spoken term in a large volume of speech segments. In order to provide fast search, speech segments are first indexed into an intermediate representation using speech recognition engines which provide multiple hypotheses for each speech segment. Approximate matching techniques are usually applied at the search stage to compensate the poor performance of automatic speech recognition engines during indexing. Recently, using visual information in addition to audio information has been shown to improve phone recognition performance, particularly in noisy environments. In this paper, we will make use of visual information in the form of lip movements of the speaker in indexing stage and will investigate its effect on STD performance. Particularly, we will investigate if gains in phone recognition accuracy will carry through the approximate matching stage to provide similar gains in the final audio-visual STD system over a traditional audio only approach. We will also investigate the effect of using visual information on STD performance in different noise environments.

Index Terms: Spoken term detection, keyword spotting, audio visual phone recognition, DMLS system

1. Introduction

STD is the process of finding all occurrences of a specified search term in a large volume of speech database [1]. This process usually consists of two steps: indexing and search. In the indexing stage, audio segments are transcribed into an intermediate representation and in the next stage, this representation is searched to detect the query terms. The indexing stage plays an important role in STD. However, it is prone to errors due to errors introduced in recognition engines used in the indexing process.

A common approach to indexing audio segments is to perform word-based transcription by using a large vocabulary continuous speech recognition (LVCSR) system [2] to produce word lattices for each audio segment [3]. The main disadvantage of such systems is that they are only able to recognize the words within their dictionary. Sub-word based strategies have been investigated to provide open vocabulary query search. The dynamic match lattice spotting (DMLS) technique [4] has been proposed as a phonetic STD approach to search and detect query terms in recognized lattices of audio segments, created using a phone recognition engine based on hidden Markov models (HMM). This technique has continued to be used as one of the state-of-the-art approaches for STD [5, 6].

Usually, in a speech document, there are other sources of information which can be used to improve the performance of indexing and search. Previously, we used topic information of the indexed documents to improve the performance of STD sys-

tem [7, 8]. Visual information is another source which has been proved to improve speech recognition performance [9, 10, 11]. Liu et al. [12] suggested early integration strategy for combining audio and visual information for the task of word spotting. The results of the experiments in this study showed that in noisy conditions, the audio visual keyword spotting system outperforms the audio-only system. In another attempt, Liu et al. [13] proposed an audio visual keyword spotting system for a robot. Obviously, robots should be able to work in real-world situations and an important aspect of such systems which should be dealt with is the presence of noise. Their experiments showed that the decision fusion of the scores of the audio and visual speech recognition models for keyword spotting significantly improves the noise robustness and provides better performance than feature fusion based audio visual spotter. Liu et al. [14] also proposed an audio visual keyword spotting system for Mandarin language and claimed that their keyword spotting system is robust to noise.

Although these approaches have shown that audio-visual fusion is superior than acoustic keyword spotting, their set of search terms were specified at indexing stage, since they needed to model the keywords before search stage which does not provide open vocabulary search. On the other hand, they could not improve the search accuracy in clean conditions. In this research we exploit DMLS approach for our baseline system and will investigate the effect of using visual information in the indexing stage on STD performance.

2. DMLS system

The phonetic STD system developed by Wallace et al. [15] which is based on the DMLS system [4] is used as our baseline system. In this system, indexing is run once to create a database from recognized lattices of phonemes and in the search phase, this database is explored to find the best match with query terms.

2.1. Indexing

The purpose of indexing is to construct a database that provides fast search. First, phonetic speech recognition is performed to decode each speech segment in the database which results in producing lattices of multiple phone sequence recognition hypotheses. In the next step, these lattices are traversed by means of Viterbi dynamic search method to extract all phone sequences with a predefined fix length, N , that terminate at each node in the lattice. All these phone sequences are then collected into a SDB. This database can be considered as a look-up table that returns the location of each occurrence of a particular unique N -gram phone sequence. In this paper, we used the value of $N = 11$, which provides a reasonable trade-off between index size and simple retrieval of long phone sequences [16].

10.21437/Interspeech.2015-203

2.2. Search

In the search stage, the query term is decomposed into its phoneme constituents using a pronunciation lexicon. Letter to sound rules are applied in case of out-of-vocabulary search terms. The difference between the target phone sequence and each indexed phone sequence is calculated using the minimum edit distance (MED) criteria. The insertion, deletion, and substitution costs for MED are trained using phone confusion network in the indexing stage. If the difference is lower than a pre-specified threshold value, then the putative occurrence is emitted as a detected occurrence.

The MED is defined as the minimum possible sum of the costs of phone substitution, insertion and deletion errors that transform the indexed phone sequence into the target phone sequence, and is calculated using dynamic programming [4]. Costs of phone substitution, insertion and deletion errors are learned from tuning data by using the confusion matrix from the recognized lattices.

3. Proposed audio visual STD system

In this paper, we propose incorporating visual information in addition to audio information in the indexing stage. We will use two approaches to train our audio visual phone HMMs to be used in the indexing stage. The first method is joint training for state synchronous HMMs [10] which is the most extensively used approach to train audio visual HMMs. The second approach is cross-database training [17] which uses data from different databases in fused HMM (FHMM) adaptation framework [18].

3.1. Joint training

In a single-stream HMM, in order to model the generation of a sequence of speech feature vectors $\{o_t\}$ of dimensionality D with the c^{th} HMM, the emission (class conditional observation) probabilities are modelled by Gaussian mixture densities,

$$Pr[o_t | c] = \sum_{k=1}^{K_c} w_{ck} \mathcal{N}_D(o_t; m_{ck}, s_{ck}) \quad (1)$$

and the HMM transition probabilities between the various classes are given by $r = [\{Pr[c' | c''], c', c'' \in C\}]^T$. Therefore, the HMM parameter vector which should be learned from training data would be,

$$a = [r^T, b^T]^T, \text{ where } b = [\{w_{ck}, m_{ck}^T, s_{ck}^T\}^T, k = \{1, 2, \dots, K\}, c \in C]^T \quad (2)$$

in which, $c \in C$ denotes the HMM states, w_{ck} are positive weights of each mixture adding to one within each state, K is the number of mixtures, and $\mathcal{N}_D(o; m, s)$ is the D -variate normal distribution with mean m and a diagonal covariance matrix, whose diagonal being represented as s .

In an audio-visual synchronous HMM, observation emission probability in each state is defined as,

$$Pr[o_t^{(AV)} | c] = \prod_{s \in \{A, V\}} \left[\sum_{k=1}^{K_{sc}} w_{sck} \mathcal{N}_{D_s}(o_t^s; m_{sck}, s_{sck}) \right]^{\lambda_{sct}}, \quad (3)$$

where all of the parameters are the same as single stream HMM parameters in Equation 1, except that λ_{sct} denotes the stream

exponents (weights) and the audio-visual feature vector is defined as,

$$o_t^{(AV)} = [o_t^{(A)T}, o_t^{(V)T}]^T. \quad (4)$$

According to Equation 3, parameters of synchronous HMM can be defined as,

$$a = [r_{av}^T, b_a^T, b_v^T, \alpha]^T, \quad (5)$$

where r_{av}^T is the state transition probability, b_a^T , and b_v^T denotes the observation emission probability of each stream, and α is the weight parameter of each stream. Training a synchronous HMM is the process of learning these parameters using standard Baum-Welch technique from a set of training data. The stream weight parameter is typically set by maximising recognition performance on a tuning dataset.

3.2. Cross-database training

In order to jointly train an audio-visual HMM, a fairly large and annotated audio-visual database is needed to learn the optimum HMM parameters. However, due to capturing difficulties, annotation cost, and time limitations, currently there are not many publicly available and annotated audio-visual databases which cover all of speech events in day-to-day conversations.

Dean et al [19] proposed continuous FHMM for audio-visual speech recognition in which a two-step framework for training audio visual models is suggested where each step could be performed on independent data. Recently, we [17] deployed this framework to use different databases to improve phone recognition performance. Cross-database training of FHMM adaptation thus consists of the following steps:

- Train an audio HMM for each phone on an external large audio database.
- For each audio observation in the given audio-visual database, find the best sequence of states of the audio HMM corresponding to that audio segment by forced-alignment.
- Train a global background GMM using all visual feature vectors of the given audio-visual database.
- For each state of each model, adapt the background model to the corresponding visual feature vectors resulted from step 2 and append it to that state as the visual GMM.

We perform phone recognition using phone models trained by these two methods and perform indexing in DMLS system to create phone sequence database. We also perform indexing using only audio data to compare audio visual STD with audio-only STD approach. The same search process will be applied on the produced database.

4. Experiments and results

4.1. Evaluation

STD can be evaluated in two stages. The first stage can test the performance of phonetic lattices using HTK style phone recognition accuracy [10]. The second stage, which actually tests the whole STD performance, is evaluated in this paper using figure of merit (FOM). FOM is used widely to report the performance of STD systems [15, 20, 21] and is defined as the average detection rate at each integer value between 0 and 10 false alarms per search term per hour.

Table 1: The configurations used for training, tuning, and test

Configurations		
Train	F02, F04, F06, F08 F10, F11, M01, M03 (CID)	F02, F04, F06, F08 F10, F11, M01, M03 (TIMIT)
	F03, F07, M02 (CID)	F03, F07, M02 (TIMIT)
Tune	1) F05, F09, M04 (CID)	1) F05, F09, M04 (TIMIT)
	2) F05, F09, M04 (TIMIT)	2) F05, F09, M04 (CID)

4.2. Training and testing datasets

Training, tuning, and testing data were extracted from the audio-visual database of spoken American English (AVDBAE) [22]. There are 14 speakers in this database including 10 females (F02-F011) and 4 males (M01-M04). Each participant reads 238 different words and 166 different sentences. The spoken text were drawn from the following sources:

1. Central Institute for the Deaf (CID) Everyday Sentences
2. Northwestern University Auditory Test No. 6
3. Vowels in /hVd/ context (separate words)
4. Texas Instruments/Massachusetts Institute for Technology (TIMIT) sentences

In this work, we only used the CID and TIMIT sentences which are quite longer than the other utterances. These sentences were divided into different portions to be used for training, tuning, and testing purposes. We also defined two configurations for our experiments: in the first configuration, CID sentences are used for training and tuning and in the second configuration, TIMIT sentences are used for training and tuning. We report the testing result (phone recognition and STD accuracy) on both sets of TIMIT and CID sentences. This was done to investigate the effect of using different and the same set of sentences in training and testing. Table 1 summarizes the portions of the dataset utilised and the different configurations used in this work. For training external audio models, TIMIT, Wall Street Journal-1, and 160 hours of speech from Switchboard-1 Release 2 was used. These models were then used for cross-database training of audio visual models.

Phone recognition is reported by calculating the phone recognition accuracy on all test segments. For STD evaluation, a total of 26 6-phone and 8-phone terms are chosen randomly from a pool of words that occur at least once in the test data to calculate FOM.

4.3. Feature extraction

Perceptual linear prediction (PLP) based cepstral features were extracted to represent the acoustic features in our experiments. Each feature vector consisted of the first 13 PLPs including the zeroth, as well as the first and second time derivatives of those 13 features. These 39 dimensional feature vectors were extracted from every 10 milliseconds of 25-millisecond windows using Hamming-windowed speech signals.

In order to extract visual features, the Fourier Lucas-Kanade algorithm proposed by Lucey et. al [23] was used to extract the lip region-of-interest (ROI) from 29.97 fps video data. This method has been shown to work better than the Viola-Jones algorithm in a semi automatic manner [24]. After ROI extraction, the mean ROI over the video segment is removed. In the

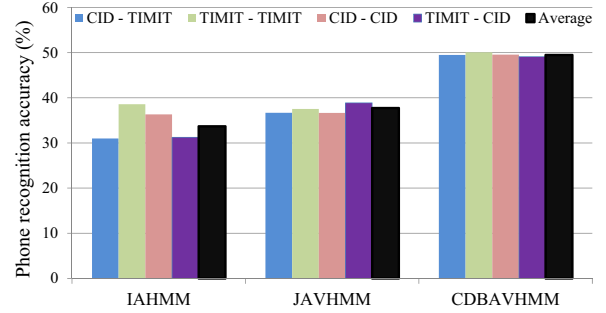


Figure 1: Phone recognition accuracy for different systems using different train/test configurations.

next step, a two-dimensional discrete cosine transform (DCT) is applied to the mean-removed ROI, with the 100 top DCT coefficients according to the zig-zag pattern retained, resulting in a ‘static’ visual feature vector. Finally, in order to extract dynamic speech information, 7 neighbouring adjacent feature vectors centred at the current vector were passed to inter-frame linear discriminant analysis (LDA) to achieve a 60 dimensional LDA feature vector.

4.4. Audio visual STD in clean environments

HMM parameters including the number of states and mixtures, grammar scale, insertion penalty, number of tokens, as well as stream weights in case of audio-visual HMMs were tuned on the tuning set after training. For audio visual models, the weight of each modality is also tuned on the tuning set. The best set of tuned parameters were then selected to report the accuracy on the test set.

Audio models (IAHMM system) are trained on internal audio data of the training set of the given audio visual database. Audio visual models which are trained using joint training method (JAVHMM system) are trained on audio visual data of the given audio visual database and HMM parameters are tuned on audio visual data of the tuning set. Audio visual models which are trained using cross-database training method (CDBAVHMM system) are first trained on audio data of the external audio database and then adapted on visual data of the training set of the given audio visual dataset and HMM parameters are tuned on audio visual data of the tuning set.

Phone recognition accuracy is reported in Figure 1. As the figure suggests, the average of phone recognition accuracy is improved when using both audio and visual modalities using joint training method compared with audio only approach. Particularly, when training and tuning on a set of sentences different than the set of test sentences (CID-TIMIT and TIMIT-CID), audio visual models are shown to have greater improvements and this suggests that adding visual information generalizes audio models to unseen speech events. As previous studies showed [17], using large external audio databases and adapting them to visual observations of the given audio visual database which is shown in CDBAVHMM system improves the phone recognition performance. This approach provides a more general audio model which could be fused with visual data and improve the performance of phone recognition. The average phone recognition accuracy of JAVHMM system is 12.0% relative higher than that of IAHMM system. This improvement is 46.9% for CDBAVHMM system. Now it will be investigated if this improvement is repeated for STD performance.

Table 2: The average of the best STD performances across different test sets for different systems

	IAHMM	JAVHMM	CDBAVHMM
FOM	0.223	0.24	0.567

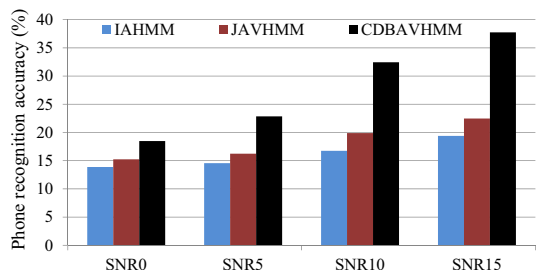


Figure 2: Phone recognition accuracy for different systems in different noise situations.

Table 2 denotes the average STD performance of the three mentioned systems in terms of FOM metric. The results shows that both audio visual systems outperform the audio only STD system. This suggests that phone recognition improvement in the indexing stage is translated into STD performance improvement when adding visual information in the form of lip movements of the speakers.

4.5. Audio visual STD in noisy environments

The QUT-NOISE [25] dataset is designed for performance evaluation of speech processing algorithms across different levels of noise recordings. In this dataset 5 noise scenarios were considered, where for each scenario, two locations were used for recording environment noise. Two sessions were conducted for recording the noise in each location which resulted in total of 20 noise recordings. In this paper we used HOME-LIVING scenario. The reason is that one of the most likely applications of audio visual STD is to search for spoken words in video chats which usually take place in HOME-LIVING area. We added the noise recording of the first session to tuning set for tuning DMLS system parameters and the second session to test set for evaluation purposes. Noise recordings are added to the original recordings at SNR levels of 0, 5, 10, and 15.

The average of the best phone recognition accuracy of each system over all test configurations for each noise level is reported in Figure 2. As it was expected, visual information provides improvements when the audio source is noisy. In all 4 noise levels, we can see that the performance of audio visual phone recognition system is better than the IAHMM system.

Figure 3 and 4 reports the results of STD experiments using the the mentioned systems in the indexing stage. As the figures show, by increasing the weight of the audio stream from 0.0 (which represents the visual-only system) in audio visual models, the accuracy of both STD systems increases until it reaches an optimum point. After that, the performance gets degraded until the audio weight is set to 1.0, which represents the audio only STD performance that is less than the optimum performance. It clearly shows that in all noise conditions, visual modality is helpful for the task of STD.

As the figures shows, for JAVHMM system, the best audio weight for high noise levels (SNR0) is at 0.6, while this value is equal to 0.8 for low levels of noise. The same be-

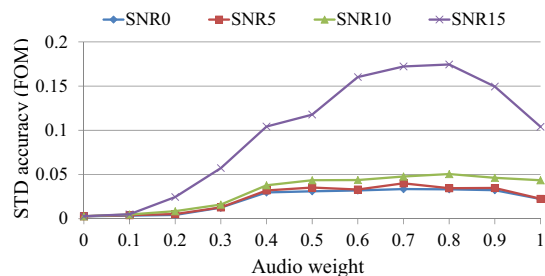


Figure 3: STD accuracy (FOM) of JAVHMM system in different noise situations.

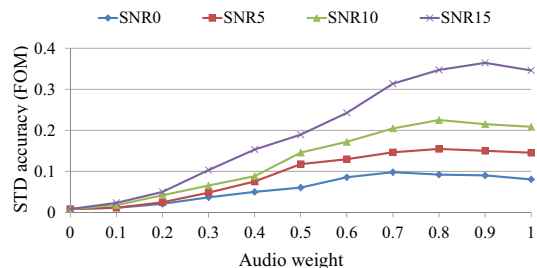


Figure 4: STD accuracy (FOM) of CDBAVHMM system in different noise situations.

haviour is observed for CDBAVHMM. This means that environments with high noise get more benefit from incorporating visual modality in STD. However, comparing both systems at the same time, we can see that the best performance for JAVHMM system is generally achieved with higher audio weight. For example, in SNR=15, the best performance of CDBAVHMM system is achieved when the audio weight is set to 0.9 while for JAVHMM it is set to 0.8. This suggests that the audio models created in CDBAVHMM system are more powerful than those of JAVHMM and therefore, it shows that visual modality is more helpful when less amount of audio data is available. The best STD performance belongs to CDBAVHMM system which was the most accurate system in the indexing stage.

5. Conclusion

In this paper, we proposed using visual information in addition to audio information for open vocabulary STD. We performed joint training of HMMs to incorporate visual information in speech models. As an alternative, the recently proposed cross-database training method was used for multimodal speech modelling. Through a set of experiments, we showed that as it was shown before, the phone recognition accuracy is increased by audio visual modelling compared with audio only speech models. We then performed STD evaluation to investigate the effect of visual information in STD performance. Results showed that visual information is indeed useful for STD. It was also shown that using visual information provides phone recognition and STD improvements in noisy environment over the audio only system.

6. Acknowledgements

This work has been supported by the Australian Cooperative Research Center for Smart Services.

7. References

- [1] National Institute of Standards and Technology, "The Spoken Term Detection (STD) 2006 evaluation plan," September 2006. [Online]. Available: <http://www.itl.nist.gov/iad/mig/tests/std/2006/>
- [2] S. Young, "A review of large-vocabulary continuous-speech," *Signal Processing Magazine, IEEE*, vol. 13, no. 5, pp. 45–, Sept 1996.
- [3] D. R. H. Miller, M. Kleber, C.-L. Kao, O. Kimball, T. Colthurst, S. A. Lowe, R. M. Schwartz, and H. Gish, "Rapid and accurate spoken term detection," in *INTERSPEECH'07*, 2007, pp. 314–317.
- [4] K. Thambiratnam and S. Sridharan, "Rapid yet accurate speech indexing using Dynamic Match Lattice Spotting," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, no. 1, pp. 346–357, 2007.
- [5] R. Wallace, B. Baker, R. Vogt, and S. Sridharan, "The effect of language models on phonetic decoding for spoken term detection," in *ACM Multimedia Workshop on Searching Spontaneous Conversational Speech*, 2009, pp. 31–36.
- [6] M. Rajabzadeh, S. Tabibian, A. Akbari, and B. Nasersharif, "Improved dynamic match phone lattice search using viterbi scores and Jaro Winkler distance for keyword spotting system," in *Artificial Intelligence and Signal Processing (AISP), 2012 16th CSI International Symposium on*, 2012, pp. 423–427.
- [7] S. Kalantari, D. Dean, and S. Sridharan, "Topic dependent language modelling for spoken term detection," in *European Signal Processing Conference, 2014. Proceedings. (EUSIPCO 2014)*, 2014. [Online]. Available: <http://eprints.qut.edu.au/75760/>
- [8] —, "Phonetic spoken term search using topic information," in *Science and Speech Technology conference*, 2014.
- [9] L. Rabiner and B.-H. Juang, *Fundamentals of Speech Recognition*. Upper Saddle River, NJ, USA: Prentice-Hall, Inc., 1993.
- [10] S. Young, G. Everman, T. Hain, D. Kershaw, G. Moore, J. Odell, V. V. Ollason, D. D. Povey, and P. Woodland, "The htk book (for htk version 3.2.1)," 2002.
- [11] G. Potamianos, C. Neti, G. Gravier, A. Garg, and A. Senior, "Recent advances in the automatic recognition of audio-visual speech," *Proceedings of the IEEE*, vol. 91, no. 9, pp. 1306–1326, Sept 2003.
- [12] M. Liu, Z. Xiong, S. Chu, Z. Zhang, and T. Huang, "Audio visual word spotting," in *Acoustics, Speech, and Signal Processing, 2004. Proceedings. (ICASSP '04). IEEE International Conference on*, vol. 3, May 2004, pp. iii–785–8 vol.3.
- [13] H. Liu, T. Fan, and P. Wu, "Audio-visual keyword spotting based on adaptive decision fusion under noisy conditions for human-robot interaction," in *Robotics and Automation (ICRA), 2014 IEEE International Conference on*, May 2014, pp. 6644–6651.
- [14] —, "Audio-visual keyword spotting for mandarin based on discriminative local spatial-temporal descriptors," in *Pattern Recognition (ICPR), 2014 22nd International Conference on*, Aug 2014, pp. 785–790.
- [15] R. Wallace, R. Vogt, and S. Sridharan, "Spoken term detection using fast phonetic decoding," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2009, pp. 4881–4884.
- [16] —, "A phonetic search approach to the 2006 NIST Spoken Term Detection evaluation," in *Interspeech*, 2007, pp. 2385–2388.
- [17] S. Kalantari, D. Dean, and S. Sridharan, "Cross database training of audio-visual hidden Markov models for phone recognition," 2015.
- [18] H. Pan, S. Levinson, T. Huang, and Z.-P. Liang, "A fused hidden Markov model with application to bimodal speech processing," *Signal Processing, IEEE Transactions on*, vol. 52, no. 3, pp. 573–581, March 2004.
- [19] D. B. Dean, P. J. Lucey, S. Sridharan, and T. J. Wark, "Fused HMM-adaptation of multi-stream HMMs for audio-visual speech recognition," in *8th Annual Conference of the International Speech Communication Association (Interspeech 2007)*. Antwerp: International Speech Communication Association (ISCA), 2007, pp. 666–669. [Online]. Available: <http://eprints.qut.edu.au/13351/>
- [20] R. G. Wallace, "Fast and accurate phonetic spoken term detection," Ph.D. dissertation, Queensland University of Technology, 2010. [Online]. Available: <http://eprints.qut.edu.au/39610/>
- [21] A. J. K. Thambiratnam and S. Sridharan, "Dynamic match lattice spotting for indexing speech content," U.S. Patent 11/377 327, August 2, 2007.
- [22] S. Richie, Carolyn Warburton and M. Carter, "Audiovisual database of spoken American English," *Linguistic Data Consortium*, 2009.
- [23] S. Lucey, R. Navarathna, A. B. Ashraf, and S. Sridharan, "Fourier lucas-kanade algorithm," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 35, no. 6, pp. 1383–1396, June 2013.
- [24] 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., Ternélie resort Le Pré du Lac, Annecy, France, 2013. [Online]. Available: <http://eprints.qut.edu.au/62749/>
- [25] D. B. Dean, S. Sridharan, R. J. Vogt, and M. W. Mason, "The qut-noise-timit corpus for the evaluation of voice activity detection algorithms," in *Interspeech 2010*, Makuhari Messe International Convention Complex, Makuhari, Japan, September 2010, to download the full database, visit: <https://wiki.qut.edu.au/display/saivt/QUT-NOISE-TIMIT>. [Online]. Available: <http://eprints.qut.edu.au/38144/>