UCBN: A new audio-visual broadcast news corpus for multimodal speaker verification studies

Girija Chetty and Michael Wagner

Human Computer Communication Laboratory,
School of Information Sciences and Engineering
University of Canberra, Australia
girija.chetty@canberra.edu.au

Abstract

The performance of face, voice, and multimodal speaker verification systems in complex and non-controlled scenarios, is typically lower than systems developed in highly controlled environments. With the aim to facilitate the development of robust multi-modal speaker recognition systems, a new multi-modal (audio-visual) Australian broadcast UCBN (University of Canberra Broadcast News) corpus was developed by capturing about 30 hours of television daily news program from several free-to-air Australian TV channels over a period of two years. In this paper we describe the acquisition of UCBN, and a new video preprocessing technique used for detection of newscasters and anchor person shots in news sequences. The speaker verification experiments using feature fusion of acoustic and visual speech features extracted from the mouth region are also reported. The performance of the complex UCBN database is compared with that of the controlled VidTIMIT database.

1. Introduction

With biometrics applications becoming more widely accepted and commonplace in high security situations, such as airports and banks, it is not hard to imagine a time in the near future where they will become ubiquitous in every day life. Less critical applications such as online identification (ID) for low cost E-commerce transactions, such as on-line shopping and video rental for example, will emerge; with the emphasis focusing on convenience and cost rather than false acceptances. Traditional problems with background noise and random lighting will become increasingly critical, because, operational conditions will not be as controlled as in high security applications. To overcome some of these problems, the use of multimodal biometric systems operating under uncontrolled environments will become more widespread.

However, there is a shortage of large multi-modal databases in the research community that can address these critical real world conditions. The M2VTS audiovisual database (Pigeon and Vandendorpe, 1997) was recorded when audio-visual speech processing was in its infancy. It consists of recordings of 37 subjects counting ten digits in French, with five recording sessions per subject, spaced weekly. Only one sentence was recorded per session. The recordings conditions were "ideal" with controlled lighting. The extended XM2VTS database addressed some of the limitations of the M2VTS database, in terms of the number of subjects and sentences. It consists of 295 subjects and the recording of three sentences. The main problem encountered with the XM2VTS database was the extremely well controlled visual recording conditions and the blue screen background. This does not represent a practical real world scenario. The recent BANCA database (Bailliere, Bengio, Bimbot, Hamouz, Kittler, Mariethoz et al, 2003) is a large audio-visual database consisting of 208 subjects, recorded under controlled, degraded and adverse scenarios. The 208 subjects are split equally into four different languages groups. This database provides data of a more challenging and practical nature for the testing of audio-visual fusion methodologies.

The UCBN database is a new audio-visual **U**niversity of **C**anberra **B**roadcast **N**ews database, consisting of five recording sessions of 30 subjects over a period of one year. The database is designed to be realistic and challenging with all of the sessions captured from complex indoor and outdoor scenarios such as noisy "real world" outdoor anchor person shot sequences with no control on illumination or background acoustic noise, and news cast sequences with complex audio-visual backgrounds. Consisting mainly of text independent (TI) recordings and limited text dependent recordings, it is cost-effective and suitable for the development of robust audio-visual speaker verification techniques. This paper reports the development of UCBN, and is organised as follows. Section 2 describes the challenges of using audio-visual broadcast news corpora for speaker verification studies, and need of better video processing in terms of speaker segmentation, shot activity detection, transcription and annotation. Section 3 and 4 describe the new audio and video pre-processing and segmentation technique developed, and details of transcription, and annotation of UCBN corpus. The

Proceedings of the 11th Australian International Conference on Speech Science & Technology, ed. Paul Warren & Catherine I. Watson. ISBN 0 9581946 2 9
University of Auckland, New Zealand. December 6-8, 2006.

Accepted after full paper review

application of the developed database for speaker verification experiments is described in rest of the sections.

2. Broadcast news corpus challenges

In previous speaker recognition studies, the evaluation of audio-video (AV) corpus design and associated technology development was not done to the same extent as the design for audio-only corpora. Millar, Wagner and Goecke in 2004 highlighted an important perspective for corpus design namely, "targeted" design versus "opportunistic" design. The aim of targeted design methodology is facilitate a particular set of experiments to answer a specific question or support a specific application. Opportunistic design, on the other hand, aims at the utilization of existing data, such as television broadcasts, and facilitates useful scientific experimentation by careful selection of already available data.

The UCBN corpus implements "opportunistic" corpus design technique, based on recording several free-to-air news broadcasts, and provides realistic and challenging scenarios for investigating robust speaker verification techniques. Though, UCBN development based on opportunistic corpus design technique does not offer the same degree of data definition as a targeted corpora, with a sound design criteria, and flexibility in terms of adaptation to suit the scientific or developmental purpose at hand, makes the opportunistic corpus design an extremely time and cost-efficient technique of collecting required data.

The opportunistic UCBN corpus design, however, in general, poses a large number of challenges. In Canberra, there are currently 5 free-to-air TV stations with 25 daily news broadcasts, a local cable television provider carries a further 7 international channels with hourly news programs around the clock, and a 3-metre satellite dish delivers more than 30 additional stations with regular news programs. For speaker verification research for example, it is possible to record from these news programmes, with between 4 and 10 newsreaders per station per week and between 5 and 50 sessions per newsreader per week.

Due to huge amount of audio visual data available from broadcast news source, audio and video pre-processing, and indexing plays a key role. The main objective of the indexing process for instance, would be to assign labels to the audio visual data in order to describe its content. The data explosion problem can however be alleviated if, before using brute force analysis tools on the entire video sequence, the relevant parts of the sequence are detected. In particular, we want to locate those parts where the audio corresponds to the face (if any) present in the image. Based on preliminary experiments, we have estimated that for more than 80% of the time in broadcast news, the audio and video do not match with respect to who is speaking, while in the remaining 20% the voice and face match, which justifies the interest of this work.

The UCBN database in current state consists of audio, video and annotation transcripts of about 30 hours of television daily news program from Australian free-to-air broadcast news channels. Next section describes the details of transcription and annotation for UCBN.

3. Transcription and Annotation of UCBN database

The transcription and annotation work involved manual as well as semi-automatic transcription with *Transana*, an open source transcription and annotation tool developed by University of Wisconsin that allows researchers to transcribe and analyze large collections of video and audio data. With Transana, we viewed the video clip first, created a transcript, and linked places in the transcript to frames in the video. We then identified and organized analytically interesting portions of video, and attached keywords to those video clips. This provided a mechanism for searching and extracting text dependent portions in the video by keywords and by combinations of keywords. The database and file manipulation tools of Transana was used for organization and storage of news casts into text dependent and text independent sections. The tool embeds automatic time codes during the transcription process. Figure 1 shows different windows in the transcription tool.

Figure 1: Transcription tool used for UCBN

3.1. UCBN database

3.1.1. Data collection

The UCBN database consists of several long daily news TV shows provided by the Australian broadcast channels. We decided to capture 5-7PM daily news programs and some mid-day broadcasts from channels ABC, SBS, Channel 7, 9, and 10, which are some of the most popular daily news programs in Australia according to TV ratings. The program provides different type of international, national and local news and special broadcasts dedicated to sports, financial, cultural news and weather reports. Table 1 and Figure 2 the recorded broadcasts by type and time duration of the news.

Figure 2: Examples of face images in UCBN corpus (speaker ID shown in Table 2)

Proceedings of the 11th Australian International Conference on Speech Science & Technology, ed. Paul Warren & Catherine I. Watson. ISBN 0 9581946 2 9
University of Auckland, New Zealand. December 6-8, 2006.

Accepted after full paper review

Table 1: Speaker IDs for UCBN faces in Figure 2 (left to right)

Speaker ID	TV Channel	Program
MS1	SBS	6.30 pm world news
MS2	ABC	7.00 pm world news
FS1	ABC	7.00 pm world news
MS3	ABC	7.30 pm current affairs
FS2	SBS	6.30 pm world news
FS3	SBS	6.30 pm world news
FS4	ABC	7.00 pm world news
FS5	SBS	6.30 pm world news
MS4	ABC	7.00 pm world news
FS5	ABC	7.00 pm world news

We had collected 30 different broadcasts from May 2004 to August 2006 . They were captured using Hauppage digital terrestrial receiver. The recordings comprised audio, video and some teletext data. Audio data were sampled at a 16000 Hz sample rate and stored as 16-bit PCM encoded mono waveform audio files. Video files were compressed and stored in MPEG2 Video format. Teletext data includes subtitling text captured simultaneously during recording of the broadcast news, however not used in any of the experiments in this study.

3.1.2. UCBN Content Analysis

To evaluate the usefulness of data for speaker verification experiments, we analyzed the content richness and consistency of annotations in UCBN. Total time duration of the collected broadcast news data is around 36 hours from which 30 hours corresponds to newscasts and reports, 3 hours of fillers (around 10% of transcribed data) and around 4 hours of data belongs to jingles and commercial breaks, which was not transcribed. The acoustic variability of newscasts was spread over a wide range. About 42% is baseline planned speech, and 20% is spontaneous. A considerable amount of around 10% of speech had background music, due to the fact that almost all headline news in most of the fillers and broadcast shows with cultural news have music in the background. A significant proportion of material (around 24%) was also degraded acoustically, and a small proportion (~ 2%) in miscellaneous conditions such as speech from non-native speakers.

Another considerable issue concerning acoustic variability is the number and distribution of speakers represented in the database. The total number of speakers were around 100. The database includes mainly native speakers with few non-native and foreign-language speakers, with 20 male and 20 female speakers producing around 40% of the speech material. There were 5 male and 5 female anchor speakers each of whom produced approximately 1hour of speech.

The analysis of linguistic content of the corpus showed that UCBN corpus consists of more than 50 different topics and 20 different filler sections. Reports covering local news represents 60% of data, international news 10%, sport news 15%, finance and business news 4%, cultural news topics 4%, and weather reports 7% of all data in the corpus. This statistics is based on time duration of typical half hour 6 pm news casts.

4. Speaker Segmentation

As broadcast news sequences are usually composed of the following elements:

- News headlines, with or without the reporters.
- Shots where the news anchor or the reporters are speaking and they are either in the scene or narrating another scene.
- News stories where either the anchor or reporter is narrating the scene or the person that is the subject of the story. The goal of automatic speaker segmentation is to detect shots that contain audio and video cuts.

The speaker segmentation technique involved both audio and video segmentation. Audio segmentation technique based on segmentation of continuous audio stream is described next.

4.1.1. Audio Segmentation

The audio segmentation locates all the boundaries between speakers in the audio signal. Some audio based speaker segmentation systems are based on silence detection. These systems rely on assumption that utterances of different people are separated by significant silences. However reliable systems would require co-operative speakers which are not the case in broadcast news. We used the ΔBIC algorithm used for partitioning the UCBN audio data based on Bayesian Information Criterion (BIC), (Delacourt and Wellekens, 2000).

Δ*BIC* is a two-pass segmentation technique. In the first pass distance between adjacent windows is obtained every 100 ms yielding in a distance signal *d(t).* Several distance measures such as symmetric Kullback-Leibler distance (KL2), Bhattacharya distance (BHA), and Arithmetic harmonic sphericity distance measures can be used, however symmetric Kullback-Leibler distance measure (Siegler, Jain & Stern, 1996), was used for implementation here. The significant peaks of *d(t)* are considered as turn candidates. In the second pass, the turn candidates are validated using Δ*BIC* criteria. To that end, the acoustic vectors of adjacent segments are modeled separately using Gaussian models. The model of the union of the acoustic vectors of both segments is also computed and then the Δ*BIC* criteria is used to check if the likelihood of the union is greater than the likelihood of both the segments individually. If likelihood of the union is greater then the turn point is discarded. Otherwise, the turn point is validated.

4.1.2. Video Segmentation

Several approaches have been used in past for detection of cuts in video sequences. Most of them rely on the similarity of consecutive frames. We developed a similarity measurement technique called the *average absolute frame difference* (AAFD) given as:

Proceedings of the 11th Australian International Conference on Speech Science & Technology, ed. Paul Warren & Catherine I. Watson. ISBN 0 9581946 2 9
University of Auckland, New Zealand. December 6-8, 2006.

Accepted after full paper review

$$AAFD(n) = \frac{1}{IJ}\sum_{i=1}^{I}\sum_{j=1}^{J}\left| f_n(i,j) - f_{n-1}(i,j)\right| \qquad (1)$$

Where I and *J* are the horizontal and vertical dimensions of the frames, n is the frame index, and *(i, j)* are the spatial co-ordinates.

The causes of dissimilarity between consecutive frames include: Actual scene transition, motion of objects, camera motion, luminance variations (flicker). In standard (good condition), the last contribution is normally negligible (except for special situations such as the presence of flash lights). Motion of objects and camera normally occur during more than one transition which produces wide pulses in AAFD. On the other hand, an abrupt scene transition produces a peak of width one in AAFD. This difference can be exploited to distinguish motion and cuts in video. Basic morphological operations, such as openings and closings, can be applied for this purpose. The proposed algorithm for cut detection can be summarized as follows:

- Obtain AAFD(n)
- Compute the residue of the morphological opening of AAFD (n).
- Threshold the residue to locate the cuts.

Many sophisticated video cut detection algorithms have been proposed in literature. However, this simple algorithm provides reasonably good results, since other transition effects such as wipes and fades are not usually used in newscasts and anchor person reports. Moreover, the anchor person reports do not usually have a high shot activity (usually the scene is an anchor person); therefore the false alarm rate within these intervals is nearly zero.

4.1.3. Audio and Video Correspondence

Once the audio and video segments are located the objective is to find the correspondence between then. Figure 3.a shows the ideal situation that we are trying to find, i.e. the audio and video segments overlap. However, for real sequences the borders of audio and video segments do not overlap, as shown in figure 3.b. This is mainly because silence periods are usually located in the audio segment borders creating a small inaccuracy. Figure 4.c shows an example of the typical situation for report segments, where a long audio segment coexists with short video segments. Given an audio segment in the time interval $[t_{min1}, t_{min2}]$ and a video segment defined in the interval $[t_{min1}, t_{min2}]$.
The intersection interval is defined as:

$$[t_{min\cap}, t_{max\cap}] = [\max(t_{min1}, t_{min2}), \min(t_{max1}, t_{max2})]$$

Then if $(t_{max\cap} - t_{min\cap}) > 0$ for a pair of audio and video segments, we define the overlap degree as:

$$overlap = \min\left\{\frac{(t_{max\cap} - t_{min\cap})}{(t_{max1} - t_{min1})}, \frac{(t_{max\cap} - t_{min\cap})}{(t_{max2} - t_{min2})}\right\}$$

If the overlap > 0.9 then the audio and video segments are said to match and a new index entry into the database is created.

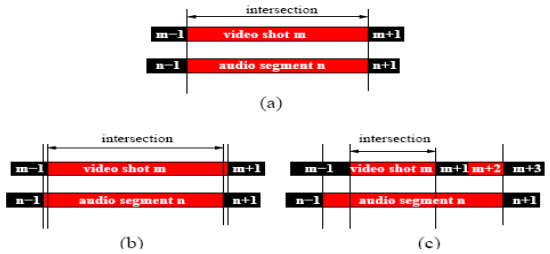

Figure 3: (a) Audio and video borders match exactly, (b) Audio and video borders almost match, (c) Audio segment contains several video shots

4.1.4. Shot activity detection

Shots where the voice corresponds to the talking face are characterized by a low activity since usually the camera is focussing a person who is talking and standing somewhere. This assumption can be used to discard some of the selected shots in the previous section. A simple activity measure is used for shot detection. The measure can be computed without full decoding of the MPEG video stream, and allows fast video segmentation the shot activity measure is based on AAFD computed in previous section to locate the video cuts. The proposed shot activity descriptor is the average value of AAFD (n) in the shot:

$$SA_1 = \sum_{n=ns}^{ne} \frac{AAFD\ (n)}{(ne - ns + 1)}$$

Where ne and ns are the initial and final frame number of the analysed shot. Figure 4 shows the frame activity for 60 seconds broadcast news using SA_1 descriptor.

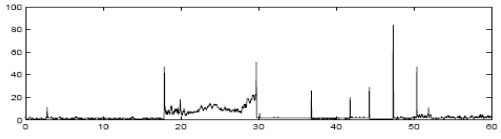

Figure 4: Shot detection using AAFD

5. UCBN for Multimodal Speaker Verification

The developed UCBN corpus was employed for multimodal (audiovisual) speaker verification experiments, where, audiovisual refers to the fusion of acoustic and visual speech. Visual speech based speaker recognition differs from well known face recognition problem in two major ways. Firstly, face recognition employs the entire face area whereas visual speech based speaker recognition employs a region of interest about the speaker's mouth, where most of the speech information is contained. Secondly, for face recognition, a gallery of static face images forms a template, whereas for visual speech based speaker recognition, visual speech temporal characteristics are used for modeling the speaker. The visual speech signal is rarely used as a complete recognition system; rather it is integrated with the acoustic

Proceedings of the 11th Australian International Conference on Speech Science & Technology, ed. Paul Warren & Catherine I. Watson. ISBN 0 9581946 2 9
University of Auckland, New Zealand. December 6-8, 2006.

Accepted after full paper review

speech signal. Visual speech based speaker verification is closely related to speech reading (visual speech recognition).

Recently, several state of the art reviews have been carried out on audio-visual speech processing (Potamianos and Neti, 2003; Chibelushi and Deravi, 2002). It has been shown in Chetty and Wagner (2004a), that feature fusion is a better strategy than classical late fusion for visual speech based speaker verification, particularly to thwart fraudulent replay attacks. Feature fusion involves extracting audio and visual features from the entire utterance, and uses concatenated (combined) audio-feature vectors for building the stochastic (GMM) speaker models; whereas, late fusion involves attaining separate audio and visual scores and combining those using appropriately chosen weights. Indeed, several studies have reported visual speech based speaker recognition results separately (Luettin, Thacker and Beet, 1996; Wark, Sridha and Chandran, 2000), with surprisingly low error rates. In Luettin et al. (1996), an accuracy of 2.1% was reported based on shape and intensity features, however, on a subject set of just 12 speakers. In Wark et al.(2000) a lip based HMM attained a speaker ID error rate of 16% on the larger M2VTS database. It can be argued that these examples of high performance were based on unrealistic scenarios (controlled recordings), and that in a more realistic environment the performance will be significantly lower. To test the effect of this, speaker verification experiments using mouth region visual speech features were carried out for both UCBN and controlled VidTIMIT corpus Sanderson and Paliwal, 2003).

The VidTIMIT database consists of video and corresponding audio recordings of 43 people, reciting short sentences. The database was recorded in 3 sessions, with a mean delay of 7 days between Session 1 and 2, and 6 days between Session 2 and 3. The sentences were chosen from the test section of the TIMIT corpus. There are 10 sentences per person. The first six sentences (sorted alpha-numerically by filename) are assigned to Session 1. The next two sentences are assigned to Session 2 with the remaining two to Session 3. The first two sentences for all persons are the same, with the remaining eight generally different for each person. The recording was done in an office environment using a broadcast quality digital video camera. The video of each person is stored as a numbered sequence of JPEG images with a resolution of 512 x 384 pixels. 90% quality setting was used during the creation of the JPEG images. The corresponding audio is stored as a mono, 16-bit, 32 kHz WAV file. Figure 6 shows some face images from VidTIMIT database.

Figure 5: Sample face images in VidTIMIT corpus

6. Experiments Results and Discussion

In this section performance of speaker segmentation based on audio and video pre-processing for UCBN corpus is discussed first, followed by discussion on comparative performance of UCBN with respect to VidTIMIT for visual speaker identity verification task.

6.1. Speaker Segmentation Performance

The audio and video pre-processing algorithms for speaker segmentation were tested on several news sequences. The results were evaluated with following parameters:

$$DR = 100 \times \frac{num\ automatic\ \text{detections}}{num\ actual\ anchor\ persons}$$

$$FAR = \frac{num\ false\ alarms}{num\ act\ persons + num\ false\ alarm}$$

and, Selected Time *(ST)*

$$ST = \frac{total\ \ duration\ \ of\ \ selected\ \ shots}{Sequence\ \ duration}$$

Without taking into account shot activity, the DR was 95%, FAR was 42% and ST was 30%, while there is an improvement of 10 % – 15 % in results with inclusion of shot activity descriptor. It can be seen how the algorithm allows discarding a large portion of the sequence from consideration with minimal processing. Almost all false detected shots correspond to anchor person shots where the speaker is a reporter.

6.2. Speaker Verification Performance

Speaker identity verification performance based on visual speech was carried out on both VidTIMIT and UCBN databases. Utterances of around 5 second duration were used for building text independent GMM speaker models. 10 subjects from VidTIMIT were tested and compared with 10 subjects in UCBN. Two sessions were used for training and one for testing.

The ROI for visual speech feature extraction included a square window about the centre of the speaker's mouth. The mouth centre was determined automatically for every 10th image frame and interpolated for the other frames based on technique proposed in (Chetty and Wagner., 2004b). The speaker-camera distance for the VidTIMIT recordings was reasonably constant and hence scaling of the face is unnecessary. The speaker-camera distance for the UCBN varies significantly; hence, face scale normalization was required. The speaker's inter eye center pixel distance is scaled for each video frame so that it is constant. The extracted 52 x 52 pixel ROI is down sampled to 40 x 40

Proceedings of the 11th Australian International Conference on Speech Science & Technology, ed. Paul Warren & Catherine I. Watson. ISBN 0 9581946 2 9
University of Auckland, New Zealand. December 6-8, 2006.

Accepted after full paper review

pixels in order to reduce the data dimensions involved and to allow efficient implementation of the image transforms. To compensate for varying illumination, the gray scale of ROI is histogram equalised followed by mean pixel value subtraction. The image transforms are then applied to the pre-processed ROI. Figure 6 compares the ROI for two subjects of the UCBN with two subjects from VidTIMIT. The variation in illumination/scale across the different UCBN sessions is evident, whereas VidTIMIT exhibits a high level of illumination/scale consistency across different sessions; however this does not represent a real world scenario.

(a) (b)

Figure 6: Visual Speech ROI for UCBN and VidTIMIT corpus

For all experiments, feature fusion of acoustic mel frequency cepstral coefficients (MFCCs) with visual speech features (DCT/PCA) extracted from mouth ROI vector was used

It was found that 8-10 Gaussian mixture speaker model gives best performance. The equal error rates achieved for speaker ID experiments for the static (S) and static delta-acceleration (SDA) features is discussed here. The best VidTIMIT EERs achieved is of the order of 6.37%, and is almost comparable to UCBN with best EERs of 7.17%. This highlights the effect on performance due to the complex and uncontrolled video recording conditions in UCBN, inspite of use of efficient segmentation and pre-processing technique.

Also, the performance deteriorates faster with respect to feature dimension (the L highest static features) for the clean data (VidTIMIT) compared to the noisy data (VALID) due to difficulty in learning audio-visual correspondences for longer SDA feature vectors. It should be noted that higher feature dimensions result in increased computational time and that an L value of 20, resulting in a SDA dimension of 60, would be computationally too expensive to implement practically. As expected, the PCA features outperform the DCT features for all of the tests. However, the PCA features give only a 1.2% relative performance increase for the VidTIMIT SDA feature test with L=20; and a larger 5.9% increase for the corresponding UCBN test. This suggests that for clean data the DCT features do not significantly outperform the PCA features, yet, for noisier data, the DCT features may be more important. The use of delta/acceleration features does not yield a significant improvement over the static features alone (VidTIMIT: SDA 6.37% versus S 6.14% and for UCBN: SDA 7.17% versus S 7.02%).

7. Conclusions and Summary

A new multi-modal (audio-visual) broadcast news database UCBN has been created for speaker ID verification studies. The need for audio and video pre-processing and segmentation for useful data extraction from Broadcast news corpus is highlighted in this study. Also results on this database was compared with a controlled VidTIMIT corpus.

8. References

Bailliere, E. B., Bengio, S., Bimbot, F., Hamouz, M., Kittler, J., Mariethoz, J., Matas, K., Messer, V., Popovici, F., Poree, B., Ruiz, B., & Thiran, J. P.(2003) , The BANCA Database and Evaluation Protocol, in Proc. Audio- and Video-Based Biometric Person Authentication, AVBPA, Guildford, UK, 625-638.

Chetty G., and Wagner M.(2004a), Liveness Verification in Audio-Video Speaker Authentication, *in Proc. 10th ASSTA conference*, (pp. 358-363) Macquarie University Press, Australia.

Chetty G., and Wagner M. (2004b), Automated lip feature extraction for face-voice person authentication Speaker Authentication, *in Proc. IVCNZ04 conference*, (pp. 17-21), University of Canterbury Press, New Zealand.

Chibelushi C.C., Deravi F., and Mason J. S. D.(2002), A Review of Speech-Based Bimodal Recognition, *IEEE Transactions on Multimedia, vol. 4*, 23-35.

Delacourt, P., & Wellekens, C. J., (2000), DISTBIC: A speaker based segmentation for audio data indexing. *Speech Comm., Volume 32, September 2000, 1:* 111–126.

Federico, M., Giordano, D., Coletti, P., (2000), Development and Evaluation of an Italian Broadcast News Corpus, in *Proc. of the LREC 2000, Greece, Volume 2, 2:* 921–924.

Luettin, J., Thacker, N. A., and Beet, S. W.,(1996) Speaker Identification by Lipreading, i*n Proc Fourth International Conference on Spoken Language, ICSLP 96, vol.1*, 62-65.

Millar, J., Wagner, M., and Goecke, R., (2004) Aspects of Speaking-Face Data Corpus Design Methodology, *in Proc. 8th International Conference on Spoken Language Processing ICSLP2004*, 1157-1160.

Pigeon S., and Vandendorpe L., (1997), The M2VTS Multimodal Face Database (Release 1.00), *in Proc. First International Conf. on Audio- and Video-based Biometric Person Authentication, Switzerland*, 403-409.

Potamianos, G., Neti, C., Gravier, G., Garg, A., and Senior A. W.,(2003) Recent Advances in the Automatic Recognition of Audiovisual Speech, *in Proc. IEEE, vol. 91*, 1306-1324.

Sanderson, C., and Paliwal K.K., (2003), Fast features for face authentication under illumination direction changes, *Pattern Recognition Letters 24*, 2409-2419.

Siegler,M.A., Jain, U., Raj, B., & Stern R.M.,(1996)., Automatic segmentation, classification and clustering of broadcast news audio, *in Proc. Ninth Spoken Language Systems Technology Workshop*, New York.

Wark, T., Sridharan, S., and Chandran, V.,(2000), The use of temporal speech and lip information for multi-modal speaker identification via multi-stream HMMs, *in Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing,. ICASSP '00. vol. 6*, 2389-2392.

Proceedings of the 11th Australian International Conference on Speech Science & Technology, ed. Paul Warren & Catherine I. Watson. ISBN 0 9581946 2 9
University of Auckland, New Zealand. December 6-8, 2006.

Accepted after full paper review