# Annotation of Heterogeneous Multimedia Content Using Automatic Speech Recognition

Marijn Huijbregts, Roeland Ordelman, and Franciska de Jong

University of Twente, Dept. of Electrical Engineering, Mathematics and Computer Science,
P.O. Box 217, 7500 AE, Enschede, The Netherlands
{m.a.h.huijbregts,ordelman,fdejong}@ewi.utwente.nl
http://hmi.ewi.utwente.nl

**Abstract.** This paper reports on the setup and evaluation of robust speech recognition system parts, geared towards transcript generation for heterogeneous, real-life media collections. The system is deployed for generating speech transcripts for the NIST/TRECVID-2007 test collection, part of a Dutch real-life archive of news-related genres. Performance figures for this type of content are compared to figures for broadcast news test data.

## 1 Introduction

The exploitation of linguistic content such as transcripts generated via automatic speech recognition (ASR) can boost the accessibility of multimedia archives enormously. This effect is of course limited to video data containing textual and/or spoken content but when available, the exploitation of linguistic content for the generation of a time-coded index can help to bridge the semantic gap between media features and search needs. This is confirmed by the results of TREC series of Workshops on Video Retrieval (TRECVID)[1]. The TRECVID test collections contain not just video, but also ASR-generated transcripts of segments containing speech. Systems that do not exploit these transcripts typically do not perform as well as the systems that do incorporate speech features in their models [13], or to video content with links to related textual documents, such as subtitles and generated transcripts.

ASR supports the conceptual querying of video content and the synchronization to any kind of textual resource that is accessible, including other full-text annotation for audiovisual material[4]. The potential of ASR-based indexing has been demonstrated most successfully in the broadcast news domain. Spoken document retrieval in the American-English broadcast news (BN) domain was even declared ‘a solved problem’ based on the results of the TREC Spoken Document Retrieval (SDR) track in 1999 [7]. Partly because collecting data to train recognition models for the BN domain is relatively easy, word-error-rates (WER) below 10% are no longer exceptional[8, 9], and ASR transcripts for BN content approximate the quality of manual transcripts, at least for several languages.

In other domains than broadcast news and for many less favored languages, a similar recognition performance is usually harder to obtain due to (i) lack of domain-specific training data, and (ii) large variability in audio quality, speech characteristics and topics being addressed. However, as ASR performance of 50 % WER is regarded as a lower bound for successful retrieval, speech-based indexing for harder data remains feasible as long as the ASR performance is not below 50 % WER, and is actually a crucial enabling technology if no other means (metadata) are available to guide searching.

For 2007, the TRECVID organisers have decided to shift the focus from broadcast news video to video from a real-life archive of news-related genres such as news magazine, educational, and cultural programming. As in previous years, ASR transcripts of the data are provided as an optional information source for indexing. Apart from some English BN rushes (raw footage), the 2007 TRECVID collection consists of 400 hours of Dutch news magazine, science news, news reports, documentaries, educational programmes and archival video. The files were provided by

[1] http://trecvid.nist.gov

the Netherlands Institute for Sound and Vision[2]. (In the remainder this collection will be referred to as *Sound and Vision* data.)

This paper reports on the setup and evaluation of the speech recognition system (further referred to as SHoUT system[3] that is deployed for generating the transcripts that via NIST will be made available to the TRECVID-2007 participants. The SHoUT system is particularly geared towards transcript generation for the kind of heterogeneous, real-life media collections exemplified by the *Sound and Vision* data that will feature in the TRECVID 2007 collection. In other words, it targets adequate retrieval performance, rather than plain robust ASR.

As can be expected for a diverse content set such as the *Sound and Vision* data, the audio and speech conditions vary enormously, ranging from read speech in a studio environment to spontaneous speech under degraded acoustic conditions. Furthermore, a large variety of topics are addresses and the material dates from a broad time period. Historical items as well as contemporary video fall within the range. (The former with poorly preserved audio; latter with varying audio characteristics, some even without 'intended' sound, just noise).

To reach recognition accuracy that is acceptable for retrieval given the difficult conditions, the different parts of our ASR system must be made as robust as possible so that it is able to cope with those problems (such as mismatches between training and testing conditions) that typically emerge when technology is transferred from the lab and applied in a real life context. This is in accordance with our overall research agenda: the development of robust ASR technology that can be ported to different topic domains with a minimum effort. Only technology that complies with this last requirement can be successfully deployed for the wide range of spoken word archives that call for annotation based on speech content such as cultural heritage data (historical archives and interviews), lecture collections, meeting archives (city council sessions, corporate meetings).

The remainder of this paper is organised as follows. Section 2 provides an overview of the system structure, followed by a description of the training procedure that was applied for the SHoUT-2007a version aiming at the transcription of *Sound and Vision* data (section 3). Section 4 presents and discusses performance evaluation results of this system obtained using *Sound and Vision* data and various other available test corpora.

## 2 System structure

Figure 1 is a graphical representation of the system work flow that starts with speech activity detection (SAD) in order to filter out the audio that do not contain speech. The audio in the *Sound and Vision* collection contains all kinds of sounds such as music, sound effects or background noise with high volume (traffic, cheering audience, etc). As a speech decoder will always try to map a sound segment to a sequence of words, processing non-speech portions of the videos (i) would be a waste of processor time, (ii) introduces noise in the transcripts due to assigning word labels to non-speech fragments, and (iii) reduces speech recognition accuracy in general when the output of the first recognition run is used for acoustic model adaptation purposes.

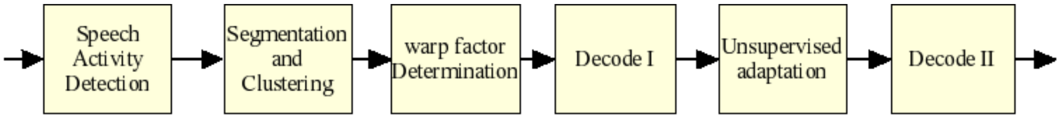
Speech Activity Detection
Segmentation and Clustering
warp factor Determination
Decode I
Unsupervised adaptation
Decode II

**Fig. 1.** Overview of the decoding system. Each step provides input for the following step.

After SAD, the speech fragments are segmented and clustered. In this step, the speech fragments are split into segments that only contain speech from one single speaker. Each segment is

[2] Netherlands Institute for Sound and Vision: http://www.beeldengeluid.nl/

[3] SHoUT is an acronym for SpeecH recognition University of Twente.

labelled with its corresponding speaker ID. Next, for each segment the vocal tract length (VTLN) warping factor is determined for vocal tract length normalisation during feature extraction for the first decoding step. Decoding is done using the HMM-based Viterbi decoder. In the first decoding iteration, triphone VTLN acoustic models and trigram language models are used. For each speaker, a first best hypothesis aligned on a phone basis is created for unsupervised acoustic model adaptation. For each file, the language model is mixed with a topic specific language model. The second decoding iteration uses the speaker adapted acoustic models and the topic specific language models to create the final first best hypothesis aligned on word basis. Also, for each segment, a word lattice is created.

The following sections describe each of the system parts in more detail.

### 2.1 Speech activity detection

Most SAD systems are based on a Hidden Markov Model (HMM) with a number of parallel states. Each state contains a Gaussian Mixture Model (GMM) that is trained on a class of sounds such as speech, silence or music. The classification is done by performing a Viterbi search using this HMM.

A disadvantage of this approach is that the GMMs need to be trained on data that matches the evaluation data. The performance of the system can drop significantly if this is not the case. Because of the large variety in the *Sound and Vision* collection it is difficult to determine a good set of data on which to train the silence and speech models. Also it is difficult to determine what kind of extra models are needed to filter out unknown audio fragments like music or sound effects and to find training data for those models.

The SAD system proposed by [1] at RT06s uses regions in the audio that have low or high energy levels to train new speech and silence models on the data that is being processed. The major advantage of this approach is that no earlier trained models are needed and therefore the system is robust for domain changes. We extended this approach for audio that contains fragments with high energy levels that are not speech.

Instead of using energy as an initial confidence measure, our system uses the output of our broadcast news SAD system. This system is only trained on silence and speech, but because energy is not used as a feature and most non-speech data will fit the more general silence model better than the speech model, most non-speech will be classified as silence. After the initial segmentation the data classified as silence is used to train a new silence model and a sound model. The silence model is trained on data with low energy levels and the sound model on data with high energy levels. After a number of training iterations, the speech model is also re-trained. The result is an HMM based SAD system with three models (speech, non-speech and silence) that are trained solely on the data under evaluation.

### 2.2 Segmentation and clustering

Similar to the SAD system, the segmentation and clustering system uses GMMs to model different classes. This time instead of trying to distinguish speech from non-speech, each GMM represents a single unique speaker. In order to find the correct number of speakers, first the data is divided into a number of small segments and for each segment a model is trained. After this, all models are pairwise compared. If two models are very similar (if it is believed that they model the same speaker) they are merged. This procedure is repeated until no two models can be found that are believed to be trained on speech of the same speaker. In [14] this *speaker diarization* algorithm is discussed in depth.

Although this speaker diarization approach has very good clustering results, it is not very fast on long audio files. In order to be able to process the entire *Sound and Vision* collection in reasonable time, we have changed the system slightly. For the purpose it is used here (to segment the data so that we can apply VTLN and do unsupervised adaptation) a less accurate clustering will still be very helpful.

The majority of processing time is spent on pairwise comparing models. At each merging iteration, all possible model combinations need to be recomputed. The longer a file is, the more initial clusters are needed. With this, the number of model combinations that need to be considered for merging increases more than linearly. We managed to bring back the number of merge calculations by simply merging multiple models at a time. The two models A and B with the highest BIC score are merged first, followed by the two models C and D with the second highest score. If this second merge involves one of the earlier merged models, for example model C is the same model as model A, also the other combination (B,D) must have a positive BIC score. This process is repeated with a maximum of four merges at one iteration. Without performance loss (see section 4), this procedure reduces the real-time factor from 8 times to 2.7 times real-time on a 35 minute TRECVID file.

### 2.3 Vocal tract length normalization

Variation of vocal tract length between speakers makes it harder to train robust acoustic models. In the SHoUT system, normalisation of the feature vectors is obtained by shifting the Mel-scale windows by a certain warping factor.

If the warping factor is smaller than one, the windows will be stretched and if the factor is bigger than one, the windows will be compressed. To normalize for the vocal tract length, large warping factors should be applied for people with a low voice and small warping factors for people with a high voice (note that this is typically because of how we implemented the warping factor. In the literature often big warping factors are linked to high voices instead of low voices).

In order to determine the speakers warping factors, a gaussian mixture model (referred to as VTLN-GMM) was trained with data from the Spoken Dutch Corpus (CGN) [10]. In total, speech of 200 male speakers and 200 female speakers was used. The GMM only contained four gaussians. For the 400 speakers in the training set, the warping factor is determined by calculating the feature vectors with a warping factor varying from 0.8 to 1.2 in step sizes of 0.04 and determining for each of these feature sets what the probability on the VTLN-GMM is. For each speaker the warping factor used to create the set of features with the highest probability is chosen.

After each speaker is assigned a warping factor, a new VTLN-GMM is trained using normalised feature vectors and again the warping factors are determined by looking at a range of factors and picking the one with the highest score. This procedure is repeated a number of times so that a VTL normalised speech model is created. From this point on, the warping factor for each speaker is determined by looking at a range of factors on this normalised model. The acoustic models needed for decoding are trained on features that are normalised using this method.

Figure 2 contains the warping factors of the hundred speakers of the VTLN-GMM training set split in male and female. Three training iterations were performed. At the third iteration a model with eight gaussians was trained. This model is picked to be the final VTLN-GMM.

### 2.4 Decoding

The decoder's Viterbi search is implemented using the token passing paradigm [6]. HMMs with three states and GMMs for its probability density functions are used to calculate acoustical likelihoods of context dependent phones. Up to 4-gram back-off language models (LMs) are used to calculate the priors.

The HMMs are organised in a single Pronunciation Prefix Tree (PPT) as described in [6, 5]. This PPT contains the pronunciation of every word from the vocabulary. Each word is defined by an ID of four bytes. Identical words with different pronunciations will receive the same ID making it possible to model pronunciation variations.

Instead of copying PPTs for each possible linguistic state (the LM n-gram history), each token contains a pointer to its LM history (see figure 3). Tokens coming from leaves of the PPT are fed back into the root node of the tree after their n-gram history is updated. Only for tokens with the same LM history, token collisions will occur. This means that each HMM state of each node

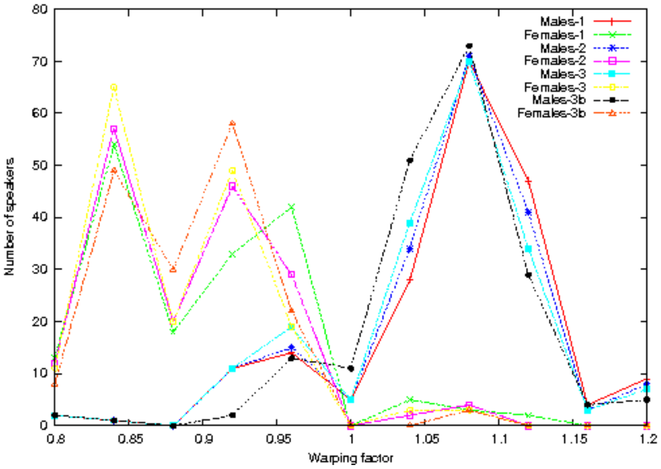

**Fig. 2.** The warping factor of 200 male and 200 female voices determined using the VTLN-GMM after each of its four training iterations. The female speakers are clustered at the left and the male speakers at the right.

in the single PPT can contain a list of tokens with unique n-gram histories. These lists are sorted in descending order of the token probability scores.

The LM data are stored in up to four lookup tables. The first table contains unigram probabilities and back-off values for all words of the lexicon. The statistics for all available bigrams, trigrams and fourgrams are stored in three minimal perfect hash tables [2]. These hash tables contain exactly one occurrence in each slot of the table. This means that no extra memory is needed except for storing the hash function and for the key of each data structure, the n-gram history. This key is needed because during lookup, the hash function will map queries for non-existing n-grams to random slots. By comparing the n-gram of the query to the n-gram of the found table slot, it can be determined if the search is successful. The algorithm proposed in [3, 2] is used to generate the hash functions. Each n-gram table contains a backoff value so that when an n-gram probability does not exist in the hash table, the system can backoff and look up the (n-1)-gram probability of the last words of the token n-gram history.

## 3 System training

### 3.1 SAD model training

The GMMs used to initialise the SAD system (see section 2.1) are trained on a small amount of Dutch broadcast news training data from the CGN corpus [10]. Three and a half hours of speech and half an hour of silence from 200 male and 200 female speakers were used. The feature vectors consist of twelve MFCC components and the zero-crossing statistic. Each GMM (silence and speech) contains 20 gaussians.

### 3.2 Acoustic model training

The acoustic models (AMs) for ASR are trained using features with twelve MFCC components. Energy is added as a thirteenth component and deltas and delta-deltas are calculated. Before calculating the MFCC features, the warping factor is determined as described in section 2.3 so that the Mel-scale windows can be shifted according to this factor.

A set of 39 triphones and one silence phone is used to model acoustics. The triphones are modelled using three HMM states. During training a decision tree is used to tie the triphone

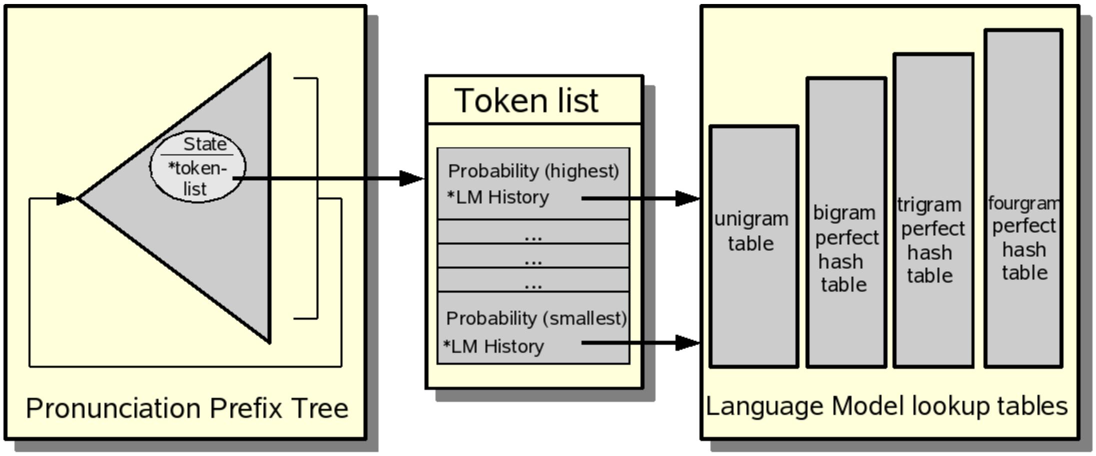
State
*token-list
Pronunciation Prefix Tree
Token list
Probability (highest)
*LM History
...
...
...
Probability (smallest)
*LM History
unigram table
bigram perfect hash table
trigram perfect hash table
fourgram perfect hash table
Language Model lookup tables

**Fig. 3.** The decoder uses a single Pronunciation Prefix Tree (PPT) and a 4-gram Language Model (LM). Each HMM state from the PPT can contain a sorted list of tokens. Each token from a list has a unique LM history.

gaussians for each state. The number of triphone clusters (triphones that are mapped to the same gaussian mixture) is limited by the amount of available training data for each cluster and a fixed maximum allowed number of clusters. After clustering the triphones, the models are trained starting with one gaussian each and iteratively increasing the number of gaussians until each cluster contains 32 gaussians.

In total the acoustic models were trained on 79 hours of broadcast news, interviews and live commentaries of sport events from the CGN corpus [10]. During training 1443 triphone clusters were defined resulting in acoustic models with in total 46176 gaussians.

This data collection is the training set that is allowed to use in the primary condition of the Dutch ASR benchmark organized by the N-Best project[4].

### 3.3 Acoustic model adaptation

The clustering information obtained during speaker diarization (see section 2.2) is used to create speaker dependent acoustic models. SMAPLR adaptation ([12]) is used to adapt the means of the acoustic model gaussians. Before adapatation starts, a tree structure is created for the adaptation data. The more data is available, the deeper the tree will be. The root of the tree contains all data and the leaves only small sets of phones. The data at each branch is used to perform Maximum a Posteriori (MAP) adaptation using output of its parent nodes as prior densities for the transformation matrices. This method prevents over-adapting models with small amounts of data but it also proves to adapt models very well if more data is available. Most important, hardly any tuning is needed using this method. Only the minimum amount of data per tree node and the maximum tree depth needed to be defined. These two parameters have been tuned on broadcast news data.

### 3.4 Language model training

Broadcast news language models (LMs) were estimated from approximately 500 million words of normalised Dutch text data from various resources (See Table 3.4). The larger part of the text data consists of written texts (98%), of which the majority is newspaper data[5]. A small portion of the available resources, some 1 million words, consists of speech transcripts, mostly derived from the various domains of the Spoken Dutch Corpus (CGN) [10] and some collected at our institute as part of the BN-speech component of the Twente News Corpus (TwNC-speech).

[4] `http://speech.tm.tno.nl/n-best`

[5] provided for research purposes by the Dutch Newspaper publisher PCM.

| corpus | description | word count |
|---|---|---|
| Spoken Dutch Corpus | various | 6.5M |
| Twente News Corpus (speech) | ’01-’02 BN transcripts | 112K |
| Dutch BN Autocues Corpus | ’98-’05 BN teleprompter | 5.7M |
| Twente News Corpus (written) | ’99-’06 Dutch NP | 513M |

**Table 1.** Dutch text corpora and word count based on normalised text.

For the selection of the vocabulary words we used the 14K most frequent words from the CGN corpus with a minimum word frequency of 10, a selection of recent words from 2006 newspaper (until 31/08/2006, see below) data and a selection of most frequent words from the 1999-2005 newspaper (NP) data set. The aim was to end up with a vocabulary of around 50K words that could be regarded as a more or less stable, general BN vocabulary that could later be expanded using task-specific words learned from either meta-data or via multi-pass speech recognition approaches. Table 2 shows that a selection of words from speech transcripts (CGN) augmented with recent newspaper data gave the best result with respect to OOV (2.33 %).

| vocabulary | %OOV |
|---|---|
| 14.6K CGN | 9.04 |
| 50K general NP (99-05) | 2.88 |
| 50K recent NP | 2.39 |
| 51K merge recent-NP, CGN | 2.33 |

**Table 2.** OOV rates using different vocabulary selection strategies based on available training data.

Trigram language models were estimated from the available data sources using modified Kneser-Ney smoothing. Perplexities of the LMs were computed using a set of manually transcribed BN shows of 7K words that date from the period after the most recent data in our text data collection. From the best performing LMs interpolated versions were created. In Table 3 perplexity results and mixture-weights of the respective models are listed. The mixture-LM with the lowest perplexity (211) was was used for decoding and contained about 8.9M 2-grams and 30M 3-grams.

| ID | LM | PP | mixture-weigths |
|---|---|---|---|
| 01 | Autocues Corpus LM | 389 | n.a. |
| 02 | CGN comp-k (news) | 797 | n.a. |
| 03 | 01 + TwNC-speech + 02 | 362 | n.a. |
| 04 | Twente News Corpus (NP) | 266 | n.a. |
| 05 | cgn-comp-f (discussion/debates) | 652 | n.a. |
| 06 | mix 03-04 | 226 | 0.44 - 0.56 |
| 07 | mix 03-04-05 | 211 | 0.27 - 0.55 - 0.18 |

**Table 3.** Language model perplexities

### 3.5 Video item specific language models

For every video item in the *Sound and Vision* collection, descriptive metadata is provided containing, among others, content summaries of around 100 words. In order to minimise the OOV rate for content words in the video items, item-specific language models were created using these descriptions, a database of newspaper articles (Twente News Corpus) and an Information Retrieval (IR) system that returns ranked lists of relevant documents given a query or ‘topic’. The procedure was as follows:

1. use description of video item from metadata as a query
2. generate LM training data from top $N$ most relevant documens given a query
3. select most frequent words
4. create topic specific language model
5. mix specific model with background language model using a fixed weight

The meta-data descriptions were normalised and stop-words were filtered out in the query processing part of the IR. For the initial experiments described in this paper, the top 50, 250 and 1000 documents from the ranked lists were taken as input for language model training. Every word from the meta-data description was added to the new vocabulary. New words from the newspaper data were only included when they exceeded a minimum frequency count (10 in the experiments reported here). Pronunciations for the new words were generated using an automatic grapheme-to-phoneme converter [11]. As no example data was available for every video item to estimate mixture weights appropriately, a fixed weight was chosen (0.8 for the NP-LM in the experiments reported here).

# 4 Experimental results

## 4.1 Evaluation data

For evaluation purposes we used a number of different resources. For evaluation of SAD and speaker diarization we used RT06s and RT07s conference meetings. These data do not match the training conditions as will also be the case with the *Sound and Vision* data and should provide a nice comparison. Also, from the *Sound and Vision* data, 13 fragments of 5 minutes each were randomly selected and annotated manually. To compare speech recognition results on *Sound and Vision* data with broadcast news transcription performance we selected one recent broadcast news show from the Twente News Corpus (TwNC-BN) that dates after the most recent date of the data that was used for language model training. For global comparison we also selected broadcast news data from the Spoken Dutch Corpus (CGN-BN).

## 4.2 Speech activity detection

In Table 4, the SAD error rate on RT06s and RT07s meeting data and *Sound and Vision* data are listed. On the RT06s and *Sound and Vision* data we also evaluated our BN SAD system (baseline). The results show that on both evaluation corpora we improved on the baseline BN SAD system. The SAD error on the conference meeting data is conform state-of-the-art[6].

| eval | baseline | SHoUT-2007a |
|---|---:|---|
| RT06s | 26.9% | 4.4% |
| RT07s | | 3.3% |
| *Sound and Vision* | 18.3% | 10.4% |

**Table 4.** Evaluation results of the SAD component showing results of the BN system (baseline) and optimised results (SHoUT-2007a).

[6] Because of the rules of the NIST evaluation we are prohibited from comparing our results with those of other participants. See `http://www.nist.gov/speech/tests/rt/rt2006/spring/pdfs/rt06s-SPKR-SAD-results-v5.pdf` for the actual ranking.

### 4.3 Speaker diarization

As we did not have ground truth transcriptions on *Sound and Vision* or BN data for diarization, we could only test diarization on conference meeting data. The speaker diarization error (DER) on the conference meetings of RT07s is 11.14% on the Multiple Distance Microphone (MDM) task. The DER on the Single Distance Microphone (SDM) task, where only one single microphone may be used which is more comparible to the task in our system, is 17.28%. Both results are conform state-of-the-art[6].

### 4.4 Automatic speech recognition

Table 5 shows the evaluation results on automatic speech recognition. It contains the Word Error Rates (WER) of the systems with and without VTLN applied. For all conditions we see a stable improvement over the baseline. It can be observed that there is a substantial performance difference between the TwNC-BN set and the CGN-BN set. This may be because the TwNC-BN data is more difficult or that the audio conditions in the CGN-BN data set better match the training conditions as a large part of our acoustic training material is derived from this corpus. Note that the CGN-BN data we used for evaluation were not in our acoustic model and language model training sets. Again it shows that *Sound and Vision* data is a difficult task domain.

| eval | baseline | SHoUT-2007a |
|---|---:|---|
| TwNC-BN | 32.8% | 28.5% |
| CGN-BN | 22.1% | 19% |
| *Sound and Vision* | 68.4% | **64.0%** |

**Table 5.** Evaluation results of the ASR component on different evaluation corpora (eval) showing Word Error Rate without VTL normalization (baseline) and with normalization (SHoUT-2007a).

### 4.5 Video item-specific language models

When we compared the baseline ASR results with the ASR approach that uses video item-specific LMs we observed that for the condition that uses the 1000 highest ranked documents for estimating the item-specific LMs some video items WER improved significantly (up to 3 % absolute) whereas for others the improvement was only marginal or even negative. On average WER improved with 0.8% absolute.

In Table 7 baseline ASR results are compared with the ASR approach that uses video item specific LMs. Only the results for the condition that uses the 1000 highest ranked documents for estimating the item specific LMs are shown as this condition produced the best results. The results indicate that using item specific LMs generally helps to improve ASR performance.

## 5 Conclusions

It is clear that ASR results on the *Sound and Vision* data leave room for improvement. When we average the results on the two types of BN data we end up with a BN-WER of 23.7% so there is a large gap between performance on BN data and performance on our target domain. The assumption is that an important part of the error is due to mismatch between training and testing conditions (speakers, recording set-ups). By implementing noise reduction techniques such as successfully applied in [1], we expect to improve system robustness on this part. On the LM level, we will fine-tune the video-specific LM algorithm and work on the lattice output of the system (rescoring with higher order n-grams).

| item | base | item-LM | delta |
|---|---|---|---|
| 01 | 73.7 | 73.5 | -0.2 |
| 02 | 55.9 | 56.3 | +0.4 |
| 03 | 53.7 | 50.7 | -3.0 |
| 04 | 74.7 | 73.8 | -0.9 |
| 05 | 79.2 | 79.4 | +0.2 |
| 06 | 42.6 | 43.2 | +0.6 |
| 07 | 60.3 | 57.9 | -2.4 |
| 08 | 62.8 | 64.1 | +1.3 |
| 09 | 59.0 | 56.6 | -2.4 |
| 10 | 39.1 | 38.7 | -0.4 |
| 11 | 74.1 | 72.4 | -1.7 |
| all | 61.4 | 60.6 | -0.8 |

**Table 6.** Results of baseline speech recognition on *Sound and Vision* data (base) compared with speech recognition results using a video item specific language models (item-LM).

| item | base | item-LM | delta |
|---|---|---|---|
| 01 | 73.7 | 73.5 | -0.2 |
| 02 | 55.9 | 56.3 | +0.4 |
| 03 | 53.7 | 50.7 | -3.0 |
| 04 | 74.7 | 73.8 | -0.9 |
| 05 | 79.2 | 79.4 | +0.2 |
| 06 | 42.6 | 43.2 | +0.6 |
| 07 | 60.3 | 57.9 | -2.4 |
| 08 | 62.8 | 64.1 | +1.3 |
| 09 | 59.0 | 56.6 | -2.4 |
| 10 | 39.1 | 38.7 | -0.4 |
| 11 | 74.1 | 72.4 | -1.7 |
| all | 61.4 | 60.6 | -0.8 |

**Table 7.** Results of baseline speech recognition on *Sound and Vision* data (base) compared with speech recognition results using a video item specific language models (item-LM).

## Acknowledgements

The work reported here was partly supported by the Dutch bsik-programme MultimediaN (`http://www.multimedian.nl`) and the EU projects MESH (IST-FP6-027685), and MediaCampaign (IST-PF6-027413). We would like to thank all those who helped us with generating speech transcripts for the *Sound and Vision* test set.

## References

1. X. Anguera, C. Wooters, and J. Pardo. Robust speaker diarization for meetings: Icsi rt06s evaluation system. In *NIST Rich Transcription 2006 Spring Meeting Recognition Evaluation, RT06s, Washington DC, USA*, volume 4299 of *Lecture Notes in Computer Science*, Berlin, October 2007. Springer Verlag.
2. A. Cardenal, J. Dieguez, and C. Garcia-Mateo. Fast lm look-ahead for large vocabulary continuous speech recognition using perfect hashing. In *proceedings ICASSP 2002*, pages 705–708, Orlando, USA, 2002.
3. Zbigniew J. Czech, George Havas, and Bohdan S. Majewski. An optimal algorithm for generating minimal perfect hash functions. *Information Processing Letters*, 43(5):257–264, 1992.
4. F. M. G. de Jong, R. J. F. Ordelman, and M. A. H. Huijbregts. Automated speech and audio analysis for semantic access to multimedia. In Y. Avrithis, Y. Kompatsiaris, S. Staab, and N. E. O'Connor, editors, *Proceedings of the First International Conference on Semantic and Digital Media Technologies, SAMT 2006, Athens, Greece*, volume 4306 of *Lecture Notes in Computer Science*, pages 226–240, Berlin, December 2006. Springer Verlag.

5. Kris Demuynck, Jacques Duchateau, Dirk Van Compernolle, and Patrick Wambacq. An efficient search space representation for large vocabulary continuous speech recognition. *Speech Commun.*, 30(1):37–53, 2000.
6. M. Finke, J. Fritsch, D. Koll, and A. Waibel. Modeling and efficient decoding of large vocabulary conversational speech. In *proceedings Eurospeech'99*, pages 467–470, Budapest, Hungary, 1999.
7. J.S. Garofolo, C.G.P. Auzanne, and E.M Voorhees. The TREC SDR Track: A Success Story. In *Eighth Text Retrieval Conference*, pages 107–129, Washington, 2000.
8. Jean-Luc Gauvain, Gilles Adda, Martine Adda-Decker, Alexandre Allauzen, Veronique Gendner, Lori Lamel, and Holger Schwenk. Where Are We in Transcribing French Broadcast News? In *InterSpeech*, Lisbon, September 2005.
9. L. Nguyen, S. Abdou, M. Afify, J. Makhoul, S. Matsoukas, R. Schwartz, B. Xiang, L. Lamel, J.L. Gauvain, G. Adda, H. Schwenk, and F. Lefevre. The 2004 BBN/LIMSI 10xRT English Broadcast News Transcription System. In *Proc. DARPA RT04*, Palisades NY, November 2004.
10. N. Oostdijk. The Spoken Dutch Corpus. Overview and first evaluation. In M. Gravilidou, G. Carayannis, S. Markantonatou, S. Piperidis, and G. Stainhaouer, editors, *Second International Conference on Language Resources and Evaluation*, volume II, pages 887–894, 2000.
11. Roeland Ordelman. *Dutch Speech Recognition in Multimedia Information Retrieval*. PhD thesis, University of Twente, The Netherlands, October 2003.
12. O. Siohan, T. Myrvol, and C. Lee. Structural maximum a posteriori linear regression for fast hmm adaptation. In *In ISCA ITRW Automatic Speech Recognition: Challenges for the Millenium*, pages 120–127, 2000.
13. Alan F. Smeaton, Paul Over, and Wessel Kraaij. Evaluation campaigns and trecvid. In *MIR 2006 - 8th ACM SIGMM International Workshop on Multimedia Information Retrieval*, 2006.
14. D. A. van Leeuwen and M. A. H. Huijbregts. The ami speaker diarization system for nist rt06s meeting data. In *NIST Rich Transcription 2006 Spring Meeting Recognition Evaluation, RT06s, Washington DC, USA*, volume 4299 of *Lecture Notes in Computer Science*, pages 371–384, Berlin, October 2007. Springer Verlag.