Audio Partitioning and Transcription for Broadcast Data Indexation

J.L. Gauvain, L. Lamel, and G. Adda

Spoken Language Processing Group
LIMSI-CNRS, BP 133, 91403 Orsay, France
{gauvain,lamel,gadda}@limsi.fr
http://www.limsi.fr/tlp

Abstract. This work addresses automatic transcription of television and radio broadcasts. Transcription of such types of data is a major step in developing automatic tools for indexation and retrieval of the vast amounts of information generated on a daily basis. Radio and television broadcasts consist of a continuous stream of data comprised of segments of different linguistic and acoustic natures, which poses challenges for transcription. Prior to word recognition, the data is partitioned into homogeneous acoustic segments. Non-speech segments are identified and removed, and the speech segments are clustered and labeled according to bandwidth and gender. The speaker-independent large vocabulary, continuous speech recognizer makes use of n-gram statistics for language modeling and of continuous density HMMs with Gaussian mixtures for acoustic modeling. The system has consistently obtained top-level performance in DARPA evaluations. An average word error of about 20% has been obtained on 700 hours of unpartitioned unrestricted American English broadcast data.

1 Introduction

With the rapid expansion of different media sources for information disemination, there is a need for automatic processing of the data. For the most part todays methods for transcription and indexation are manual, with humans reading, listening and watching, annotating topics and selecting items of interest for the user. Automation of some of these activities can allow more information sources to be covered and significantly reduce processing costs while eliminating tedious work. Radio and television broadcast shows are challenging to transcribe as they contain signal segments of various acoustic and linguistic natures. The signal may be of studio quality or have been transmitted over a telephone or other noisy channel (ie., corrupted by additive noise and nonlinear distorsions), or can contain speech over music or pure music segments. Gradual transitions between segments occur when there is background music or noise with changing volume, and abrupt changes are commonly when there is switching between speakers in different locations. The speech is produced by a wide variety of speakers: news anchors and talk show hosts, reporters in remote locations, interviews with politicians and common people, unknown speakers, new dialects, non-native speakers, etc. Speech from the same speaker may occur in different parts of the broadcast, and with different channel conditions. The linguistic style ranges from prepared speech to spontaneous speech. Acoustic models trained on clean, read speech, such as the WSJ corpus, are clearly inadequate to process such inhomogeneous data.

Two principle types of problems are encountered in transcribing broadcast news data: those relating to the varied acoustic properties of the signal, and those related to the linguistic properties of the speech. Problems associated with the acoustic signal properties are handled using appropriate signal analyses, by classifying the signal according to segment type and by training specific acoustic models for the different acoustic conditions. In order to address variability observed in the linguistic properties, we analyzed differences in read and spontaneous speech, with regard to lexical items, word and word sequence pronunciations, and the frequencies and distribution of hesitations, filler words, and respiration noises. As a result of this analysis, these phenonema were explicitly modeled in both the acoustic and language models as described in [4].

2 Data Partitioning

While it is evidently possible to transcribe the continuous stream of audio data without any prior segmentation, partitioning offers several advantages over this straight-foward solution. First, in addition to the transcription of what was said, other interesting information can be extracted such as the division

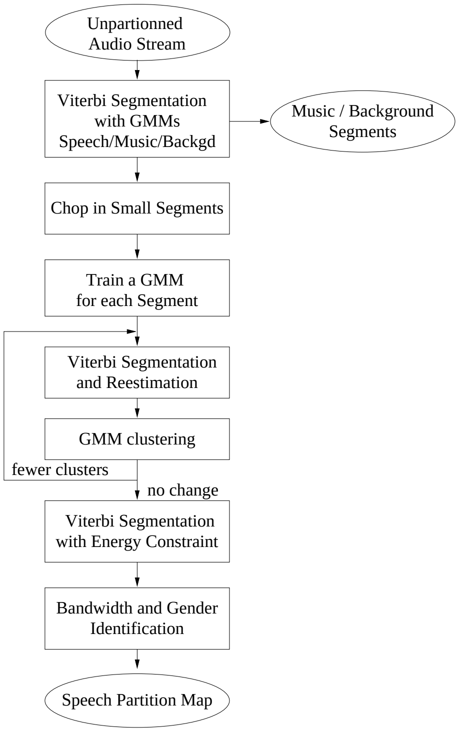
Unpartionned Audio Stream
Viterbi Segmentation with GMMs Speech/Music/Backgd
Music / Background Segments
Chop in Small Segments
Train a GMM for each Segment
Viterbi Segmentation and Reestimation
GMM clustering
fewer clusters
no change
Viterbi Segmentation with Energy Constraint
Bandwidth and Gender Identification
Speech Partition Map

Fig. 1. Partitioning algorithm.

into speaker turns and the speaker identities. Prior segmentation can avoid problems caused by linguistic discontinuity at speaker changes. By using acoustic models trained on particular acoustic conditions, overall performance can be significantly improved, particularly when cluster-based adaptation is performed. Finally by eliminating non-speech segments and dividing the data into shorter segments (which can still be several minutes long), reduces the computation time and simplifies decoding.

The segmentation and labeling procedure introduced in [5] is shown in Figure 1. First, the non-speech segments are detected (and rejected) using Gaussian mixture models. The GMMs, each with 64 Gaussians, serve to detect speech, pure-music and other (background). The acoustic feature vector used for segmentation contains 38 parameters. It is the same as the recognition feature vector except that it does not include the energy, although the delta energy parameters are included. The GMMs were each trained on about 1h of acoustic data, extracted from the training data after segmentation with the transcriptions. The speech model was trained on data of all types, with the exception of pure music segments and the silence portions of segments transcribed as speech over music. In order to detect speech in noisy conditions a second speech GMM was trained only on noisy speech segments. These model are expected to match all speech segments. The music model was trained only on portions of the data that were labeled as pure music, so as to avoid mistakenly detecting speech over music segments. The silence model was trained on the segments labeled as silence during forced alignment, after excluding silences in segments labeled as containing speech in the presence of background music. All test segments labeled as music or silence are removed prior to further processing.

A maximum likelihood segmentation/clustering iterative procedure is then applied to the speech segments using GMMs and an agglomerative clustering algorithm. Given the sequence of cepstral vectors corresponding to a show $(x_1, \ldots, x_T)$, the goal is to find the number of sources of homogeneous data (modeled by the p.d.f. $f(\cdot|\lambda_k)$ with a known number of parameters) and the places of source changes. The result of the procedure is a sequence of non-overlaping segments $(s_1, \ldots, s_N)$ with their associated segment cluster labels $(c_1, \ldots, c_N)$, where $c_i \in [1, K]$ and $K \leq N$. Each segment cluster is assumed to represent one speaker in a particular acoustic environment.

Speaker-independent GMMs corresponding to wideband speech and telephone speech (each with 64 Gaussians) are then used to label telephone segments. This is followed by segment-based gender identification, using 2 sets of GMMs with 64 Gaussians (one for each bandwidth). The result of the partitioning process is a set of speech segments with cluster, gender and telephone/wideband labels, as illustrated in Figure 2.

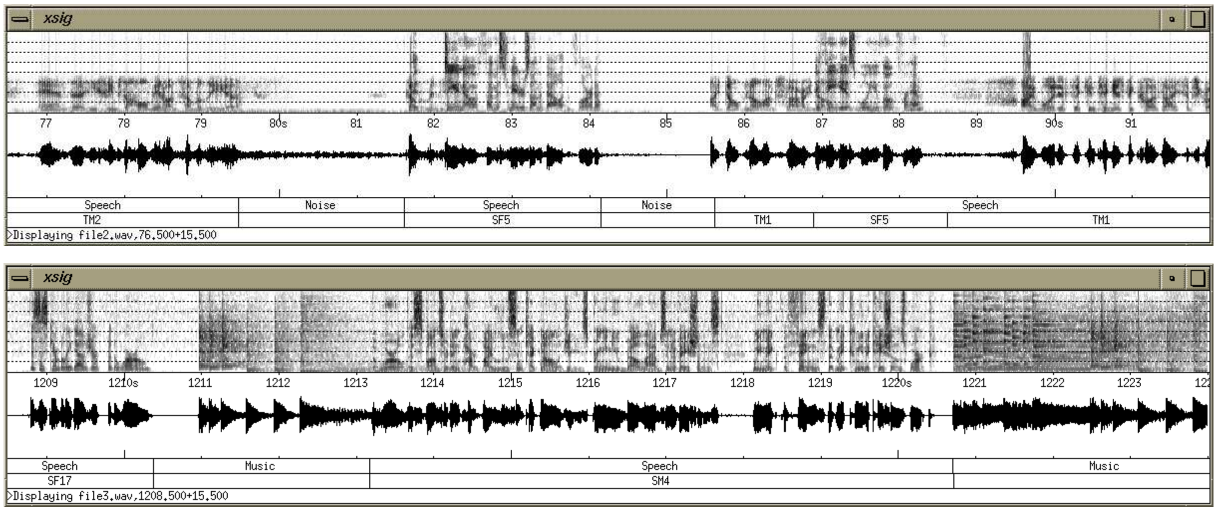

Fig. 2. Spectrograms illustrating results of data partitioning on sequences extracted from broadcasts. The upper transcript is the automatically generated segment type: Speech, Music, or Noise. The lower transcript shows the clustering results for the speech segments, after bandwidth (T=telephone-band/S=wideband) and gender (M=male/F=female) identification. The number identifies the cluster.

We evaluated the frame level segmentation error (similar to [7]) on the 4 half-hour shows in the DARPA eval96 test data using the manual segmentation found in the reference transcriptions. The NIST transcriptions of the test data contain segments that are not scored, since they contain overlapping or foreign speech, and occasionally there are small gaps between consecutive transcribed segments. Since we consider that the partitioner should also work correctly on these portions, we relabeled all excluded segments as speech, music or other background.

Show	*1*	*2*	*3*	*4*	*Avg*
Frame Error	7.9	2.3	3.3	2.3	3.7
M/F Error	0.4	0.6	0.6	2.2	1.0
#spkrs/#clusters	7/10	13/17	15/21	20/21	-
ClusterPurity	99.5	93.2	96.9	94.9	95.9
Coverage	87.6	71.0	78.0	81.1	78.7

Table 1. Top: Speech/non-speech frame segmentation error (%), using NIST labels, where missing and excluded segments were manually labeled as speech or non-speech. Bottom: Cluster purity and best cluster coverage (%).

Table 1(top) shows the segmentation frame error rate and speech/non-speech errors for the 4 shows. The average frame error is 3.7%, but is much higher for show 1 than for the others. This is due to a long and very noisy segment that was deleted. Averaged across shows the gender labeling has a 1% frame error. The bottom of Table 1 shows measures of the cluster homogeneity. The first entry gives the total number of speakers and identified clusters per file. In general there are more clusters than speakers, as a cluster can represent a speaker in a given acoustic environment. The second measure is the cluster purity, defined as the percentage of frames in the given cluster associated with the most represented speaker in the cluster. (A similar measure was proposed in [1], but at the segment level.) The table shows the weighted average cluster purities for the 4 shows. On average 96% of the data in a cluster comes from a single speaker. When clusters are impure, they tend to include speakers with similar acoustic conditions. The “best cluster” coverage is a measure of the dispersion of a given speaker’s data across clusters. We averaged the percentage of data for each speaker in the cluster which has most of his/her data. On average, 80% of the speaker’s data goes to the same cluster. In fact, this average value is a bit misleading as there is a large variance in the best cluster coverage across speakers. For most speakers the cluster coverage is close to 100%, i.e., a single cluster covers essentially all frames of their data. However, for a few speakers (for whom there is a lot of data), the speaker is covered by two or more clusters, each containing comparable amounts of data.

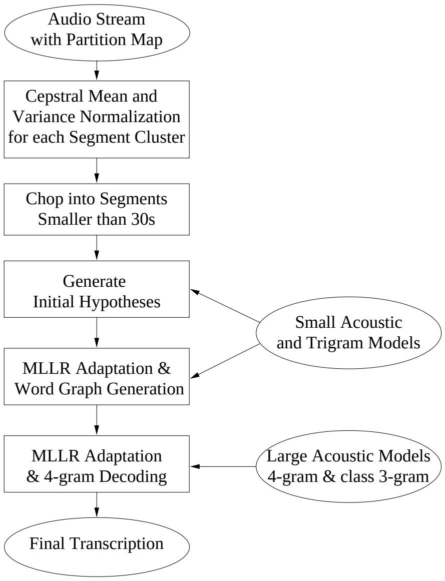

Fig. 3. Word decoding.

3 Transcribing Partitioned BN Data

The speech recognizer uses continuous density HMMs with Gaussian mixture for acoustic modeling and *n*-gram statistics estimated on large text corpora for language modeling[4]. For acoustic modeling, 39 cepstral parameters are derived from a Mel frequency spectrum estimated on the 0-8kHz band (0-3.5kHz for telephone speech models) every 10ms. The LPC-based cepstrum coefficents are normalized on a segment cluster basis using cepstral mean removal and variance normalization. Each resulting cepstral coefficient for each cluster has a zero mean and unity variance. Each context-dependent phone model is a tied-state left-to-right CD-HMM with Gaussian mixture observation densities (about 32 components) where the tied states are obtained by means of a phonemic decision tree. Gender-dependent acoustic models were built using MAP adaptation of SI seed models for wideband and telephone band speech[3]. Word recognition is performed in three steps: 1) initial hypothesis generation, 2) word graph generation, 3) final hypothesis generation, each with two passes. The initial hypothesis are used in cluster-based acoustic model adaptation using the MLLR technique[10] prior to word graph generation and in all subsequent decoding passes. The final hypothesis is generated using a 4-gram, optionally interpolated with a category trigram model with automatically generated word classes[8].

The acoustic models were trained on about 150 hours of Broadcast News data. Language models were obtained by interpolation of backoff n-gram language models trained on different data sets: BN transcriptions, NAB newspapers and AP Wordstream texts prior to Sep95 and after July96, and transcriptions of the BN acoustic training data. The interpolation coefficients of these 4 LMs were chosen so as to minimize the perplexity on the Nov96 and Nov97 evaluation test sets. The recognition vocabulary contains 65122 words and has a lexical coverage of over 99% on the Nov98 evaluation test data. The pronunciations are based on a 48 phone set (3 of them are used for silence, filler words, and breath noises). A pronunciation graph is associated with each word so as to allow for alternate pronunciations, including optional phones. Compound words for about 300 frequent word sequences subject to reduced pronunciations were included in the lexicon as well as the representation of frequent acronyms as words.

The word decoding procedure is shown in Figure 3. Prior to decoding, segments longer than 30s are chopped into smaller pieces so as to limit the memory required for the trigram decoding pass[4]. To do so a bimodal distribution is estimated by fitting a mixture of 2 Gaussians to the log-RMS power for all frames of the segment. This distribution is used to determine locations which are likely to correspond to pauses, thus being reasonable places to cut the segment. Cuts are made at the most probable pause 15s to 30s from the previous cut. Word recognition is performed in three steps: 1) initial hypothesis generation, 2) word graph generation, 3) final hypothesis generation, each with two passes.

The first step, carried out in two passes, generates initial hypotheses which are used for cluster-based acoustic model adaptation. The first pass of this step generates a word graph using a small bigram backoff language model and gender-specific sets of 5416 position-dependent triphones with about 11500 tied states. This is followed by a second decoding pass with a larger set of acoustic models (27506 triphones with 11500 tied states) and a trigram language model (about 8M trigrams and 15M bigrams) to generate the hypotheses. Band-limited acoustic models are used for the telephone speech segments.

The second step generates accurate word graphs. Unsupervised acoustic model adaptation (both means and variances) is performed for each segment cluster using the MLLR technique[10]. The mean vectors are adaptated using a single block-diagonal regression matrix, and a diagonal matrix is used to adapt the variances. Each segment is decoded first with a bigram language model and an adapted version of the small set of acoustic models, and then with a trigram language model (including 8M bigrams and 17M trigrams) and an adapted version of the larger acoustic model set.

The final hypothesis is generated using a 4-gram interpolated with a category trigram model with 270 automatically generated word classes[8]. The first pass of this step uses the large set of acoustic models adapted with the hypothesis from Step 2, and a 4-gram language model. This hypothesis is used to adapt the acoustic models prior to the final decoding step with the interpolated category trigram model.

In Table 2 reports the word recognition results on the eval test sets from the last three years. All of our system development was carried out using the eval96 data. The results shown in bold are the official NIST scores obtained by the different systems. In Nov97 our main development effort was devoted to moving from a partitioned evaluation to the unpartitioned one. The Nov97 system[5] did not use focus-condition specific acoustic models as had been used in the Nov96 system[4]. This system nevertheless achieved a performance improvement of 6% on the eval96 test data. The Nov98 system[6] has more accurate acoustic and language models, and achieves a relative word error reduction of over 20% compared to the Nov97 system.

	Test set (Word Error)		
System	*Eval96*	*Eval97*	*Eval98*
Nov96 system	**27.1***		
Nov97 system	25.3	**18.3**	
Nov98 system	19.8	13.9	**13.6**

Table 2. Summary of BN transcription word error rates. *Only the Nov96 system used a manual partition. All other results are with an automatic partition.

4 Experiments with Spoken Document Retrieval

One of the main motivations for automatic processing of the audio channels of broadcast data is to serve as a basis for automatic disclosure and indexation for information retrieval purposes. While in traditional IR tasks, the result is an ordered set of related documents, for spoken document retrieval (SDR) the result is a rank-ordered set of pointers to temporal excerpts[2]. SDR supports random access to relevant portions of audio documents, reducing the time needed to locate recordings in large multimedia databases.

We have assessed the performance in spoken document retrieval using state-of-the-art speech recognition technology. These results were obtained using hidden Markov models, with Porter stemming[12] and blind feedback, as proposed by [11]. Specifically we compare retrieval performance on automatically generated transcripts with manually produced transcripts[1] using the SDR'98 TREC-7 data. This data consists of about 100 hours of radio and television broadcasts (1997 LDC Hub4 Broadcast News corpus) and contains about 2800 stories with known boundaries. The ordered list of retrieved stories was scored using the TREC-EVAL scoring software and the NIST reference assessments. Using the automatically generated transcripts our system obtains a Mean Average Precision[2] of 0.47, which is not much less than the Mean Average Precision of 0.52 using the manual reference transcripts. An even smaller difference has been obtained using the TNO retrieval system (see [9]).

5 Summary & Discussion

In this paper we have presented our recent research in partitioning and transcribing televison and radio broadcasts. These are neccessary processing steps to enable automated processing of the vast amounts of audio and video data produced on a daily basis. The data partitioning algorithm makes use of Gaussian

[1] Here we do not consider differences in retrieval from text sources and audio materials.

mixture models and an iterative segmentation and clustering procedure. The resulting segments are labeled according to gender and bandwidth using 64-component GMMs. The speech detection frame error is less than 4%, and gender identification has a frame error of 1%. Many of the errors occur at the boundary between segments, and can involve silence segments which can be considered as with speech or non-speech without influencing transcription performance.

Word recognition is carried out in multiple passes for each speech segment using more progressively more accurate models. The generation of word graphs with adapted acoustic models is essential for obtaining word graphs with low word error rates, particularly in light of the variety of talkers and acoustic conditions. Based on our experience, it appears that current word recognition performance is not critically dependent upon the partitioning accuracy and that any reasonable approach that separates speaker turns and major acoustic boundaries is sufficient. On unrestricted broadcast news shows, such as the 1996 dev and eval data, the word error rate is about 20%.

Some preliminary experiments have been carried out to assess information retrieval performance from spoken documents. A simple HMM-based system was trained and tested using the TREC-7 data. A mean average precision of 0.47 was obtained on automatic transcripts, and 0.52 on manual reference transcripts. We are looking into ways to make use of additional information such as confidence measures to reduce this difference.

Due to the availability of large, transcribed corpora available through the LDC, our work thus far has focused on American English. In the context of the LE-4 OLIVE project we have ported our transcription system to French and a port to German is in progress. Although substantial performance improvements have been obtained over the last 3 years, there is still a need to improve the underlying speech recognition technology.

Acknowledgment

The authors thank Michele Jardino for providing the word classes used in the class-based language model.

REFERENCES

[1] S.S. Chen, P.S. Gopalakrishnan, “Speaker, Environment and Channel Change Detection and Clustering via the Bayesian Information Criterion”, *DARPA Broadcast News Transcription and Understanding Workshop*, pp. 127-132, February 1998.

[2] J.S. Garofolo, E.M. Voorhees, C.G.P. Auzanne, V.M. Stanford, B.A. Lund, “1998 TREC-7 Spoken Document Retrieval Track Overview and Results”, *Proc. 7th Text Retrieval Conference TREC-7*, 1999.

[3] J.L. Gauvain, C.H. Lee, “Maximum *a Posteriori* Estimation for Multivariate Gaussain Mixture Observation of Markov Chains,” *IEEE Trans. on SAP*, **2**(2), pp. 291-298, April 1994.

[4] J.L. Gauvain, G. Adda, L. Lamel, M. Adda-Decker, “Transcribing Broadcast News: The LIMSI Nov96 Hub4 System,” *Proc. ARPA Speech Recognition Workshop*, pp. 56-63, February 1997.

[5] J.L. Gauvain, L. Lamel, G. Adda, ”The LIMSI 1997 Hub-4E Transcription System”, *Proc. DARPA Broadcast News Transcription and Understanding Workshop*, pp. 75-79, February 1998.

[6] J.L. Gauvain, L. Lamel, G. Adda, M. Jardino, ”The LIMSI 1998 Hub-4E Transcription System”, *Proc. DARPA Broadcast News Workshop*, February 1999.

[7] T. Hain, S.E. Johnson, A. Tuerk, P.C. Woodland, S.J. Young, “Segment Generation and Clustering in the HTK Broadcast News Transcription System,” *DARPA Broadcast News Transcription and Understanding Workshop*, pp. 133-137, February 1998.

[8] M. Jardino “Multilingual stochastic n-gram class language models,” *Proc. IEEE ICASSP-96*, Atlanta, 1996.

[9] F. de Jong, J.L. Gauvain, J. den Hartog, K. Netter, “OLIVE: Speech Based Video Retrieval,” *These proceedings.*

[10] C.J. Leggetter, P.C. Woodland, “Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models,” *Computer Speech and Language*, **9**(2), pp. 171-185, 1995.

[11] D.R.H. Miller, T. Leek, R.M. Schwartz, “BBN at TREC7: Using Hidden Markov Models for Information Retrieval,” *Proc. 7th Text Retrieval Conference TREC-7*, 1999.

[12] M.F. Porter, “An Algorithm for Suffix Stripping,” *Program* **14**(3), pp. 130-137, 1980.

[13] M. Siegler, U. Jain, B. Raj, R. Stern, “Automatic Segmentation, Classification and Clustering of Broadcast News Audio,” *Proc. DARPA Speech Recognition Workshop*, Chantilly, Virginia, pp. 97-99, February 1997.