PARTITIONING AND TRANSCRIPTION OF BROADCAST NEWS DATA

Jean-Luc Gauvain, Lori Lamel, Gilles Adda

Spoken Language Processing Group
LIMSI-CNRS, BP 133
91403 Orsay cedex, FRANCE
{gauvain,lamel,gadda}@limsi.fr

ABSTRACT

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. In this paper we report on our recent work in transcribing broadcast news data[2, 4], including the problem of partitioning the data into homogeneous segments prior to word recognition. Gaussian mixture models are used to identify speech and non-speech segments. A maximum-likelihood segmentation/clustering process is then applied to the speech segments using GMMs and an agglomerative clustering algorithm. The clustered segments are then labeled according to bandwidth and gender. The recognizer is a continuous mixture density, tied-state cross-word context-dependent HMM system with a 65k trigram language model. Decoding is carried out in three passes, with a final pass incorporating cluster-based test-set MLLR adaptation. The overall word transcription error on the Nov'97 unpartitioned evaluation test data was 18.5%.

1. INTRODUCTION

In this paper we report on our recent work in transcribing broadcast news shows [2, 4], and work addressing the problem of partitioning the data into homogeneous segments for further processing. Radio and television broadcasts contain signal segments of various linguistic and acoustic natures ranging from well prepared speech of news anchors to spontaneous speech from unknown callers or interviewees. The signal may be studio quality or have been transmitted over a telephone or other noisy channel (ie., corrupted by additive noise and nonlinear distorsions), or may contain speech over music. The transition between segment types can be gradual, such as when there is background music with changing volume, or abrupt when switching between speakers in different locations. Speech from the same speaker may occur in different parts of the broadcast, and with different channel conditions. Transcription of this type of data poses challenges in dealing with the continuous stream of data under varying conditions.

When the acoustic conditions are unknown unsupervised adaptation techniques can be effective in improving performance. Such methods are more effective as the amount of adaptation data increases, therefore it is of interest to cluster segments from the same speaker and condition. The goal of data partitioning is to divide the acoustic signal into homegenous segments, and to associate appropriate labels with the segments.

2. DATA PARTITIONING

The segmentation and labeling procedure introduced in [4] is shown in Figure 1. First, the non-speech segments are detected (and rejected) using Gaussian mixture models (GMMs). Three 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 three 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. These models 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

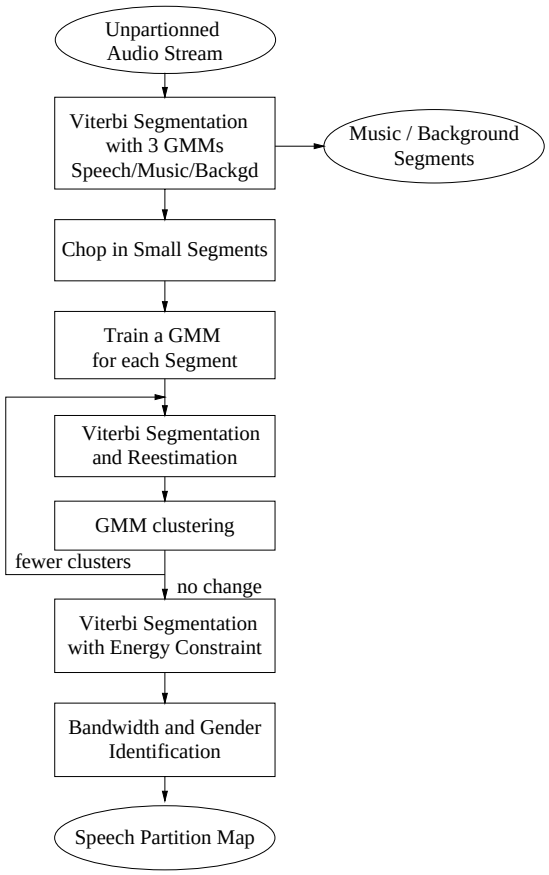

Figure 1: Partionning algorithm.

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. In absence of any prior knowledge about the stochastic process governing (K, N) and the segment lengths, we use as objective function a penalized log-likelihood of the form

$$\sum_{i=1}^{N} \log f(s_i|\lambda_{c_i}) - \alpha N - \beta K$$

where $\alpha > 0$ and $\beta > 0$. The terms αN and βK, which can be seen as segment and cluster penalties, correspond to the parameters of exponential prior distributions for N and K. It is easy to prove that starting with overestimates of N and K, alternate Viterbi reestimation and agglomerative clustering gives a sequence of estimates of (K, N, λ_k) with non decreasing values of the objective function. In the Viterbi step we reestimate (N, λ_k) so as to increase $\sum_i \log f(s_i|\lambda_{c_i}) - \alpha N$ (i.e. adding a segment penalty α in the Viterbi search) whereas in the clustering step two or more clusters can be merged as long as the resulting log-likelihood loss per merge is less than β. This algorithm stops when no merge is possible. A constraint on the cluster size is also used to ensure that each cluster corresponds to at least 10s of speech. (Recall that the previously rejected non-speech segments are not considered here.)

For single Gaussian models the merging criterion is easy to implement since the log-likelihood loss can be directly computed from the sufficient statistics of the corresponding segments[5, 7]. In the more general case of Gaussian mixtures, there are no sufficient statistics and there is no direct solution to compute the resulting mixture and/or the log-likelihood loss. We can envison estimating the new mixture from the data but this is a costly procedure. Another solution that we adoped for this work is to modify the objective function, replacing the likelihood function by the complete data likelihood of the Gaussian mixtures and extending the maximum likelihood clustering method to the Gaussian level. To estimate the log-likelihood loss for two Gaussian mixtures, we simply have to compute the sum of the log-likelihood loss while clustering the Gaussians of the 2 mixtures (until we get the desired number of Gaussians per mixture). We have used 8 mixture components per cluster, so to compute the log-likelihood loss induced by merging two clusters agglomerative clustering is performed starting with 16 Gaussians until 8 Gaussians are left.

The process is initialized using a simple segmentation algorithm based on the detection of spectral change (similar to the first step used in the CMU'96 system[8]). The threshold is set so as to over-generate segments. Initially, the cluster set consists of a cluster per segment. This is followed by Viterbi training of the set of GMMs (one 8-component GMM per cluster). This procedure is controlled by 3 parameters: the minimum cluster size (10s), the maximum log-likelihood loss for a merge (α), and the segment boundary penalty (β). When no more merges are possible, the segment boundaries are refined using the last set of GMMs and an additional relative energy-based boundary penalty, within a 1s interval. This is done to locate the segment boundaries at silence portions, so as to avoid cutting words.

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.

3. TRANSCRIBING PARTITIONED BN DATA

The word decoding procedure is shown in Figure 2. For acoustic modeling, cepstral parameters are derived from a Mel frequency spectrum estimated on the 0-8kHz band (0-3.5kHz for telephone speech models) every 10ms[2, 3]. For each 30ms frame the Mel scale power spectrum is computed, and the cubic root taken followed by an inverse Fourier transform. 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. The 39-component acoustic feature vector consists of 12 cepstrum coefficents and the log energy, along with the first and second order derivatives.

Prior to decoding, segments longer than 30s are chopped into smaller pieces so as to limit the memory required for the trigram decoding pass[2]. 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) word graph generation, 2) trigram pass, 3) cluster-based acoustic model adaptation. The word graph is generated using a 65K word bigram backoff LM. This step uses a gender-specific sets of position-dependent triphones with about 8500 tied states and a small bigram LM (about 2M bigrams). Differents acoustic models are used for telephone and wideband segments. The sentence is then decoded using the word graph generated in the first step with a large set of gender-dependent acoustic models (position-dependent triphones with about 11500 tied states) and a 65K word trigram LM (including 8M bigrams and 16M trigrams). Finally, unsupervised acoustic model adaptation (both means and variances) is performed for each segment cluster using the MLLR technique, prior to the last decoding pass with the adapted models and the trigram LM. The mean vectors are adaptated using a single block-diagonal regression matrix, and a diagonal matrix is used to adapt the variances.

Two sets of gender-dependent acoustic models have been built using MAP adaptation of SI seed models for each of wideband and telephone band speech. These models were trained on about 80 hours of transcribed broadcast news data from a variety of television and radio shows. The bigram and trigram language models were trained on newspaper texts (the 1995 Hub3 and Hub4 LM material – 155M words), on the broadcast news (BN) transcriptions (years 92-96, 125M words), and the 866K words in the transcriptions of the 95-96 acoustic training data. The BN transcriptions were processed in order to be homogeneous with the previous texts, and filler words mapped to a unique form. After transforming the training texts to be closer to the observed American reading style, they were processed in order to add a proportion of breath markers (4%), and of filler words (0.5%)[2]. Cross sentence trigram counts were added to the within sentence trigram counts before estimating the LM parameters.

The recognition vocabulary contains 65,252 words and 72,788 phone transcriptions. The vocabulary selection and language models have been optimized on the 1996 Hub-4 F0 and F1 evaluation test set. The OOV rate is 0.66% on the 1996 Hub-4 dev test data and 0.97% on F0-F1 part of the Nov'96 eval test set. Pronunciations are based on a 48

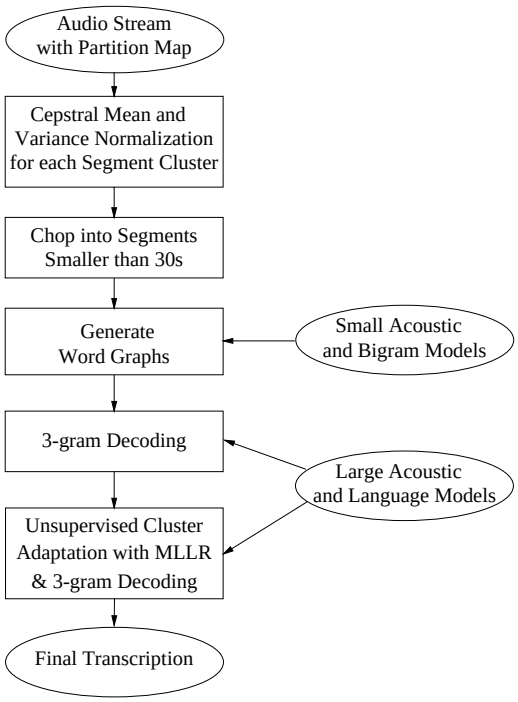

Figure 2: Word decoding.

phone set (3 of them are used for silence, filler words, and breath noises). The filler and breath phones were added to model these events, which are relatively frequent in the broadcast data and are not used in transcribing other lexical entries. A pronunciation graph is associated with each word so as to allow for alternate pronunciations, including optional phones. Frequently occuring inflected forms were verified to provide more systematic pronunciations. The lexicon contains the most common 1000 acronyms found in the training texts, and compound words to represent frequent word sequences[2]. This provides an easy way to allow for reduced pronunciations.

4. EXPERIMENTAL RESULTS

For development data we used the dev96 and eval96 data sets. In this paper we report results on the eval96 and eval97 test sets. In order to evaluate the partitioning quality, we compare the segmentation error at the frame level (similar to [6]) to the test data transcriptions for 4 half-hour shows (eval96). The NIST transcriptions of the test data contain segments that were not scored, since they contain overlapping or foreign speech, and occasionally there are small gaps between consecutive transcribed segments. We therefore relabeled all excluded segments as speech, music or other background.

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. In addition to these errors, there are 6.2%

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	8/10	12/17	14/21	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 (%).

Test set	*Corr*	*Sub*	*Del*	*Ins*	*Err*
eval96	77.8	15.4	6.9	3.1	25.3
*eval97**	84.1	12.4	3.5	2.5	18.5

Table 2: Word error rates for of unpartitioned evaluation on 1996 and 1997 eval test data. (* Official NIST score).

female speech frames deleted (largely due to the same segment) and 1.7% of the male frames deleted. 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. There are more clusters than speakers, as a cluster can represent a speaker in a given acoustic environment. We define the cluster purity to be the % of frames in the given cluster coming from 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. 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. There is a large variance in the best cluster coverage across speakers. For most speakers, a single cluster covers essentially all frames of their data. However, for some speakers for whom there is a lot of data we have observed that the speaker is covered by two clusters, with comparable amounts of data.

In Table 2 we report word recognition results on the eval96 and eval97 data sets. The high deletion rate on the eval96 data is mainly due to 2 very noisy speech segments which were classified as non-speech. (This type of error was less frequent on the eval97 data which was of higher quality on average.) However since the word error is very high on these segments, rejecting them has only a marginal effect on the overall word error rate. The result is a higher deletion rate and a lower substitution one.

5. SUMMARY

In this paper we have presented our recent research in partitioning and transcribing televison and radio broadcasts. The data partitioning algorithm makes use of Gaussian 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. Our clustering procedure tends to generate slightly more clusters than the true number of speakers in a show. The average cluster purity is 95%, with many clusters representing a single speaker. The per speaker best cluster coverages are either close to 100% or close to 50% in cases where a speaker's data was split into two equal-sized clusters.

Word recognition is carried out in multiple passes for each speech segment using more progressively more accurate models. The final decoding pass uses cluster-based test-set MLLR adaptation. The overall word transcription error of the Nov'97 unpartitioned evaluation test data (3 hours) was 18.5%. 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.

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, Feb. 1998.

[2] 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, Feb. 1997.

[3] J.L. Gauvain, G. Adda, L. Lamel, M. Adda-Decker, "Transcription of Broadcast News," *EuroSpeech'97*, pp. 907-910, Sept. 1997.

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

[5] H. Gish, M. Siu, R. Rohlicek, "Segregation of Speakers for Speech Recognition and Speaker Identification," *ICASSP-91*, pp. 873-876, May 1991.

[6] 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, Feb. 1998.

[7] A. Kannan, M. Ostendorf, J.R. Rohlicek, "Maximum Likelihood Clustering of Gaussians for Speech Recognition," *IEEE Trans. Speech and Audio*, Vol.2, no.3, July 1994.

[8] M. Siegler, U. Jain, B. Raj, R. Stern, "Automatic Segmentation, Classification and Clustering of Broadcast News Audio," *DARPA Speech Recognition Workshop*, pp. 97-99, Feb. 1997.

[9] R. Schwartz, H. Jin, F. Kubala, S. Matsoukas, "Modeling Those F-Conditions – Or Not," *DARPA Speech Recognition Workshop*, pp. 115-118, Feb. 1997.