LANDMARK BASED LARGE VOCABULARY CONTINUOUS SPEECH RECOGNITION

Daniel MORARU, Guillaume GRAVIER

Institut de Recherche en Informatique et Systemes Aleatoires,
Campus de Beaulieu, 35042 Rennes Cedex France,
Tel : +33 2 99 84 72 39 / Fax : +33 2 99 84 71 71
E-mail : daniel.moraru@gmail.com, ggravier@irisa.fr

Large vocabulary automatic speech recognition relies usually on Hidden Markov Models (HMM) which make little use of phonetic or extra-linguistic knowledge. As an alternative landmark based speech recognition relies on precise signal landmarks and exploits distinctive features. Different types of landmarks can be used: phonetic, speaker, speech type, video, etc. In this paper we will focus on two kinds of landmarks: speaker and phonetic. We propose a theoretical framework to combine both approaches by introducing prior knowledge in a non-stationary HMM based decoder. As a case study we investigate how speaker landmarks issued out of speaker segmentation can be used for speech recognition and also how broad phonetic landmarks can be integrated in a HMM based decoder in order to focus on the best search path. We will show that in this case every phonetic class brings a small improvement, the best improvement being obtained with glides. Using all broad phonetic classes brings a significant improvement by reducing the error rate from 23% to 14% on a broadcast news transcription task. We also experimentally demonstrate that landmarks do not need to be detected with precise boundaries and can be used to fasten the beam search algorithm.

Key words: Speech recognition, speaker segmentation, phonetic landmarks, hidden Markov models.

1. INTRODUCTION

In state of the art speaker recognition systems based on hidden Markov models the decoding part usually consists of building a unidirectional graph which includes all the information sources available (language model, dictionary, acoustic models) and search for the best path in this graph in order to find the sequence of words that maximises the probability:

$$\hat{w} = \arg\max_{w} \quad p(y/w)\,p(w) \tag{1}$$

using the Viterbi algorithm. The approach based on hidden Markov models is based at the acoustic level on training data and does not permit to take into consideration external phonetic or other type of landmarks. The opposite of this approach is represented by certain scientific works that try to explicitly use phonetic landmarks in order to characterize the speech signal for automatic speech recognition [1]-[6]. Those approaches rely on the detection of specific features like the onsets and the offsets of the fricatives and the voiced consonants. However, in real-life applications, the automatic detection of those kinds of features is rather difficult, especially for noisy signals and spontaneous speech.

In this paper we will try to use external *a priori* information represented by speaker related landmarks and phonetic landmarks. In the first case the speaker related information will be generated automatically using a speaker segmentation system. The speaker related information will be used for adaptation of the acoustic models. This way we should obtain specific acoustic speaker models. For the phonetic related

landmarks we propose an approach that will allow us to use them in order to guide the search for best path in the decoding graph. This way the *a priori* information is used as anchor points for the graph decoding algorithm. In this paper we try to validate our approach and to investigate the impact on the performance of a large vocabulary continuous speech recognition system. In order to correctly evaluate the performance impact, the phonetic landmarks are manually detected. However, in the proposed experimental framework the detection error can still be taken into account.

At first we will present the speaker segmentation system followed by the large vocabulary continuous speech recognition baseline system. The next two sections will concern the use of speaker and phonetic landmarks, namely the speaker adaptation approach and the use of phonetic landmarks for graph decoding algorithm. The paper will end with experiments, results and conclusions.

2. A FAST SPEAKER SEGMENTATION SYSTEM

In this section we present the IRISA speaker segmentation system. A variant of this system was successfully used during the ESTER French evaluation campaign.

The IRISA system [7] is based on a BIC (Bayesian Information Criterion) speaker change detector followed by an hierarchical clustering. The clustering stop condition is the estimation of the number of speakers using a penalized BIC criterion.

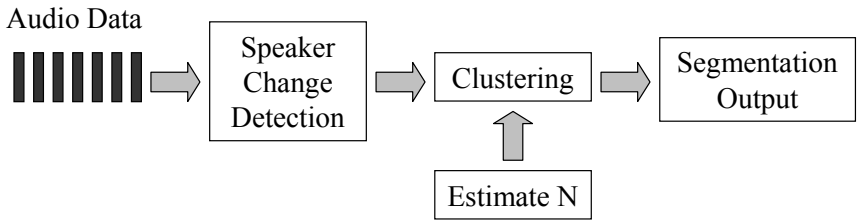

Figure 1. Audio Speaker Segmentation System.

There is a speech detection phase carried out with a 4 state ergodic hidden Markov model where the states represent speech, speech with background music, music and silence. State conditional probabilities are modeled by a 256 component mixture of Gaussians with diagonal covariance. The feature vector is a 36 component vector with 12 cepstral coefficients plus the first and second order derivatives with a mean and variance normalization. After Viterbi decoding, all segments containing speech, whether with background music or not, are tagged as speech and adjacent speech segments are merged. The miss and inserted speech rates are respectively 1.25% and 5.7% on the development set, and 1.9% and 14% on the test set of the ESTER corpus [12].

Speaker change detection is carried out independently in each of the detected speech segments with a three pass variant of the BIC change detection algorithm with 24-channel Mel filter-bank features and a full covariance matrix. The average segment length at the output of the speaker change detection algorithm is around 10 seconds, with a purity of about 9%. At the end of this phase we should obtain segments containing speech from only one speaker. Speaker must cluster those segments and the number of speakers must be estimated.

The speaker segmentation system implements a bottom-up clustering with a global BIC stop criterion, where clusters are modeled using 32 component Gaussian mixture model (GMM). Feature vectors consist of 16 mel frequency cepstral coefficients plus energy.

For a given broadcast show, clusters are first initialized from the segments by adapting the mean vectors of a generic document speech background model (DSBM) according to a MAP criterion. The DSBM parameters are estimated using the speech data from the entire show. A model-space based approximation of the Kullback-Leibler divergence between two GMMs whose mean vectors were adapted from the same model [8], is used as a distance measure between two clusters:

$$D\left(C_i, C_j\right) = \sum_k \sum_d w_k \frac{\left(m_{k,d}^i - m_{k,d}^j\right)^2}{\sigma_{k,d}^2} \tag{2}$$

where w_k is the weight of Gaussian k, $m_{k,d}$ is the element d of the mean vector of Gaussian k and $\sigma_{k,d}$ is the element d of the diagonal covariance matrix of Gaussian k.

The main advantage of this distance based solely on the parameters of two GMMs is its low computation time compared to a generalized likelihood ratio. The two closest clusters are merged until a stop criterion is reached. The clustering stop criterion is based on the detection of a global maximum of the global BIC criterion, with:

$$BIC\left(M\right) = \log L\left(X / M\right) - \lambda \frac{m}{2} \cdot NSp \cdot \log Nx \tag{3}$$

where $L\left(X / M\right)$ is the likelihood of the entire data given the current cluster models. The global BIC criterion has the advantage that the number of speakers NSp appears explicitly in the penalization term. The coefficient λ is tuned on the tuning database and applied as is on the test speech corpus.

The results for this baseline system will be presented in section 6.

3. THE SPEECH RECOGNITION SYSTEM

This speech recognition system was developed for French radio broadcast news transcription. The system is capable of recognising over 60000 word including different pronunciations of the same word. It also performs well in case of noisy conditions and bandwidth variations.

The same speech detection phase is applied as for the speaker segmentation system. Then instead of speaker change detection there is a sentence boundaries detection applied. The sentence boundaries detection consists of bandwidth segmentation followed by in-breath and out-breath group detection.

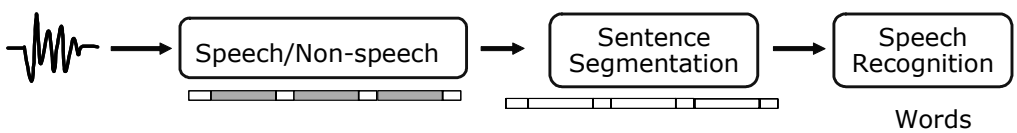

Figure 2. The Speech Recognition System

The speech recognition system is based on Hidden Markov Models decoding of speech. The speech decoding is done in two phases. On phase 1 (TRS1) we use mono-phone acoustic models characterized by 128 component GMMs. The result of this phase is a reduced word graph. On phase 2 (TRS2) this word graph is rescored using tri-phones acoustic models characterized by 32 component GMMs. In both cases the language models are tri-gram models. The whole system is based on Sirocco (local code) and HTK ToolKit [9].

The system uses beam search decoding in order to look for the most likely path in the decoding graph. This technique is based on Viterbi decoding. This algorithm is an incremental technique that looks for the partial optimal path that ends in a certain state at a certain moment $\left(j,t\right)$ using the formula:

$$S\left(j,t\right) = \max_i S\left(i,t-1\right) + \ln a_{ij} + \ln p\left(y_t / j\right) \tag{4}$$

where a_{ij} is the transition probability between the states i and j and $p\left(y_t / j\right)$ is the likelihood of emitting y_t in the state j. In this case $S\left(j,t\right)$ is the score of the best partial path leading to the state j at the moment t.

During the beam search decoding each moment t only the path with the score above a certain path are recorded and developed:

$$S(i,t) \geq \max_i S(j,t) - \alpha \tag{5}$$

where α is a selection tuning parameter which controls the beam size. We usually fix a certain number of maximum paths that will be kept.

The phonetic anchors will be used in order to limit the graph size generated by the first phase of the transcription system in order to produce smaller graphs without eliminating the true transcription path. The speaker related anchors generated by the speaker segmentation system will be used in order to adapt the acoustic models of phase 2 to every speaker detected.

4. SPEAKER ADAPTATION

In classical speaker adaptation for speech recognition we usually assume that we have extra tuning data for a certain speaker. In this case the generic acoustic models are adapted to the target speaker. Two adaptation techniques are generally used: Maximum Likelihood Linear Regression (MLLR) [10] and Maximum A Posteriori (MAP) [11].

In the case of MLLR adaptation the goal is to reduce the mismatch between an initial model set and the adaptation data. The goal is to estimate a set of linear transforms W for mean and variance using the EM (Expectation Maximisation) algorithm. In the case of linear transforms for the mean vector the adapted mean vector is obtained from the initial mean vector using:

$$\hat{\mu} = W\begin{bmatrix} 1 & \mu \end{bmatrix} \tag{6}$$

Depending on the amount of adaptation data available this type of adaptation can be applied in a very flexible manner. If a small amount of data is available then a global transform matrix can be generated. This global transform is applied to every Gaussian component in the model set. However if more adaptation data is available improved adaptation is possible by increasing the number of matrix transforms. For instance the Gaussian components could be grouped into the broad phone classes: silence, vowels, stops, glides, nasals, fricatives, etc. The adaptation data could now be used to construct more specific broad class transforms to apply to these groupings.

In HTK rather than specifying static component groupings or classes, a robust and dynamic method is used for the construction of further transformations as more adaptation data becomes available. MLLR makes use of a regression class tree to group the Gaussians in the model set, so that the set of transformations to be estimated can be chosen according to the amount and type of adaptation data that is available. The tying of each transformation across a number of mixture components makes it possible to adapt distributions for which there were no observations at all. With this process all models can be adapted and the adaptation process is dynamically refined when more adaptation data becomes available.

Model adaptation can also be accomplished using a maximum a posteriori (MAP) approach. This adaptation process is sometimes referred to as Bayesian adaptation. MAP adaptation involves the use of prior knowledge about the model parameter distribution. If we know what the parameters of the model are likely to be (before observing any adaptation data) using the prior knowledge, we might well be able to make good use of the limited adaptation data.

For MAP adaptation purposes, the informative priors that are generally used are the speaker independent model parameters. The adaptation formula for state j and mixture component m is:

$$\hat{\mu}_{jm} = \frac{N_{jm}}{N_{jm} + \tau}\bar{\mu}_{jm} + \frac{\tau}{N_{jm} + \tau}\mu_{jm} \tag{7}$$

where τ is the weighting of the prior knowledge of the adaptation speech data and N is the occupation likelihood of the adaptation data. Also, $\bar{\mu}$ is the estimated mean of the observed adaptation data. As can be

seen, if the occupation likelihood of a Gaussian component is small, then the MAP estimate will remain close to the speaker independent component mean.

In our experiments the adaptation data is obtained from the result of the speaker segmentation system. As a result the acoustic models used to transcribe the current sentence are adapted with the same speaker data obtained from different sentences. All possible transform were used and results will be presented in section 6.

5. PHONETIC LANDMARKS

We will extend the formalism of the algorithm of Viterbi in order to introduce in the process of searching the best path in the decoding graph certain constraints related to the a priori knowledge we have. For instance if we know a priori that a certain part of speech corresponds to a vowel it is possible to penalize even to suppress the paths from the decoding graph that do not respect this constraint.

A possible approach of this idea is to consider the decoding graph as being non-stationary. If a certain transition puts us in the state (i,t) which is incompatible with our constraint the cost of this transition should be increased.

In order to take into account the a priori knowledge as a constraint on the possible paths in the decoding graph, the transition probabilities a_{ij} from equation 4 should be replaced by:

$$\ln a_{ij}(t) = \ln a_{ij} - \lambda I_j(t) \tag{8}$$

where $I_j(t)$ is the constraint factor whose value is 0 if the state (j,t) of the decoding graph does not contradict the constraint. The penalty factor $\lambda \geq 0$ that is applied to a transition that is incompatible with our constraint depends on how confident are we in the a priori information we dispose. For instance, if $\lambda = \infty$, we consider our a priori information as being as reliable as possible giving them a complete confidence. In this case the search for the best path in the graph is limited to the path that respect the a priori knowledge.

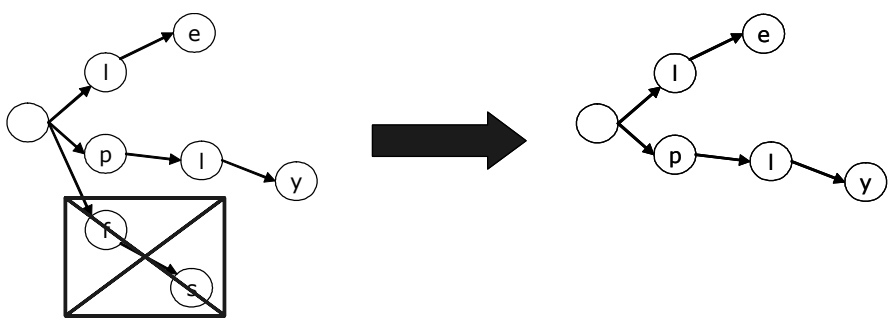

Figure 3. Graph decoding when the constraint is a vowel and $\lambda = \infty$.

If $\lambda = 0$ then we are actually doing a classical graph decoding without any constraints. For our experiments we will consider $\lambda = \infty$ as our a priori information is manually obtained and could be considered in this case as highly reliable. In our approach the a priori information is used as anchor points to guide the research of the most likely path.

In our experiments presented in section 6 we will consider 5 broad phonetic classes: vowels, plosives, fricatives, nasals and glides. The idea is to segment the signal according to those classes and keep as valid paths in the decoding graph only those that respect those constraints. Since we assume it is not possible to detect precisely the boundaries during the segmentation process, we will consider only the central part of every landmark. Since the goal of this paper is only to study if the landmark information is useful or not the landmarks will be obtain directly from manual transcription and not detected automatically. Results will be presented for every phonetic class separately and for the case where we assume that the presence of all classes is known a priori.

6. EXPERIMENTS AND RESULTS

In this section we will present the results concerning the speaker adaptation and also the landmark based speech recognition from the previous two sections. We will start however by presenting the speech corpus used for our experiments.

Our experiments were preformed on a subset of 4 files extracted from the French radio broadcast news ESTER corpus [12]. However due to the lack of time the experiments concerning the landmark based speech recognition results will be presented only on one of the files. The speech from this corpus is mostly non-spontaneous speech recorded from real radio broadcast news shows. It is usually high quality speech, sampled at 16 kHz, 16 bits. There are however some interviews with more spontaneous speech or degraded quality speech (e.g. telephone speech).

The landmarks were obtained using a forced alignment of the manual transcription of the speech corpus. For a certain phonetic class the landmarks are obtained directly from the reference phonetical alignment. For every phone the anchor is centred on it and the width of the anchor is 50 % of the phone reference duration.

The speaker landmarks were obtained using the speaker segmentation system previously presented in [7]. The speaker segmentation error for our corpus was 18.9 %. The adaptation techniques used for our experiments were MAP with different adaptation weightings and MLLR with one global transform and also MLLR with regression class tree transforms using different number of classes. The results for these experiments for MLLR with one global transform are presented in Table 1.

Table 1. Speaker adaptation results WER (%) for ESTER corpus. Speaker error for speaker segmentation system is 18.9 %

Baseline TRS1	Baseline TRS2	TRS2 MLLR
30.8 %	24.8 %	23.8 %

The reason for which only those results are presented is that the WER improvement is the same no matter which adaptation technique is used. The speaker segmentation only provides a few sentences as adaptation data for most of the speakers. The other adaptation techniques require usually more data and also more precise transcription than this obtained automatically.

The results obtained for the phonetic landmark based speech recognition are presented in Table 2 and Table 3. Table 2 presents the results in term of Word Error Rate (WER) for one of the files from the ESTER corpus. Even the results are presented only on one file the conclusions can be considered valid for the whole corpus since the files are very similar and also the results of the baseline system confirm our hypothesis.

Table 2. Phonetical landmarks results WER (%) for 20030418_0700_0800_inter_dga file only

	Baseline	All	Vowels	Plosives	Fricatives	Nasals	Glides
TRS1	29.8 %	15.1 %	27.1 %	27.0 %	28.3 %	28.2 %	26.0 %
TRS2	23.2 %	14.4 %	21.6 %	21.2 %	21.5 %	22.2 %	20.8 %

As we expected there is a lot improvement if all phonetic landmarks are used. However this improvement is equally distributed between different type of landmarks. Basically there is no class that provides a significant word error rate reduction. The most error rate reduction among the phonetic classes is provided by the glides. The improvement is visible for both speech transcription phases.

Another advantage of using the phonetic landmarks is the speed gain. The word graph produced by the first phase of the transcription is significantly reduced. In this case the graph decoding of the second phase will perform faster. Also since during the first phase not all the paths are investigated the graph construction will also be faster. Results are presented in Table 3 in terms of number of nodes and number of arcs per graph.

As expected the size of the word graph is significantly reduced if all constraints are used. In this case the size of the graph is about one third of the baseline graph. Every phonetic class helps reducing the graph without increasing the word error rate. The a priori knowledge of the glides landmarks leads as expected to the smallest graph among the individual constraints.

Table 3. Phonetical landmarks results for 20030418_0700_0800_inter_dga file only. Graph size after TRS1 expressed in number of arcs and number of nodes

	Baseline	All	Vowels	Plosives	Fricatives	Nasals	Glides
Arcs	18110	2870	13216	14499	14882	16159	12521
Nodes	2553	618	2031	2138	2182	2324	1890

7. CONCLUSIONS

In this paper we have investigated the use of external *a priori* information represented by speaker related landmarks and phonetic landmarks. In the case of speaker related landmarks the information was generated automatically using a speaker segmentation system and was used to adapt the speaker independent acoustic models.

The phonetic related landmarks were manually obtained since the goal of this study was only to investigate if this kind of information is useful for a speaker recognition system. The phonetic related landmarks were used to guide the search for best path in the decoding graph.

The results concerning the speaker related landmarks showed only a slight improvement (about 1% absolute WER) in the case were a global MLLR transform was used. This result is normal since there is not enough speaker information obtained from speaker segmentation. This makes difficult to use a more elaborated speaker related adaptation like MAP or Regression Class MLLR. Since only the central part of each phoneme was used we also showed that it is not necessary be very precise when detecting the landmarks.

The perspectives of our work concerns mainly the phonetic landmarks based speech recognition. Our first perspective would be to study the proposed approach using a real landmark detector instead of manually annotated landmarks. This would allow us to investigate the robustness of our approach with regard to the landmark time precision detection.

A second perspective would consist in trying to use an adaptive penalty value λ from equation 8 with respect to our confidence in our landmark detection system.

And finally we should consider that our approach is applicable to any problem that can be formulated as the research of the optimal path in a decoding graph and it is applicable to many domains other than speech recognition. For example this type of approach can be used for multimedia stream analysis since a certain media can provide information on other media. Those information can provide landmarks for this other media.

REFERENCES

1. CHEN M., *Nasal Landmark Detection.* International Conference on Speech and Language Processing, ICSLP 2000 Proceedings, pp. 636-639, 2000.
2. HOWITT A., *Vowel Landmark Detection.* International Conference on Speech and Language Processing, ICSLP 2000 Proceedings, 2000.
3. JOHNS HOPKINS UNIVERSITY, Center for Language and Speech Processing, *Landmark-based Speech Recognition: Report of the 2004 Johns Hopkins Summer Workshop*, 2005.
4. JUNEJA A., *Speech Recognition Based on Phonetic Features and Acoustic Landmarks.* PhD thesis, University of Maryland, 2004.
5. LI J., LEE C.-H., *On Designing and Evaluating Speech Event Detectors.* ISCA European Conference on Speech Communication and Technology, INTERSPEECH 2005 Proceedings, 2005.
6. LIU S., A., *Landmark Detection for Distinctive Feature-based Speech Recognition.* PhD. Thesis, Massachusetts Institute of Technology, 1995.
7. MORARU D., BEN M., GRAVIER G., *Experiments on Speaker Tracking and Segmentation in Radio Broadcast News Identification of Speakers Engaged in Dialog*. ISCA European Conference on Speech Communication and Technology, INTERSPEECH 2005, 2005.
8. DO M., N., *Fast Approximation of Kullback-Leibler Distance for Dependent Trees and Hidden Markov Models.* IEEE Signal Processing Letters**, Vol.10**, No. 4, pp 115-118, 2003.
9. HTK Toolkit, http://htk.eng.cam.ac.uk/
10. LEGGETTER C.,J., WOODLAND P., C., *Maximum Likelihood Linear Regression for Speaker Adaptation of Continuous Density Hidden Markov Models.* ELSEVIER Computer Speech and Language, **Vol. 9**, No. 2, pp. 171-185, 1995.

11. GAUVAIN J.-L., LEE C.-H., *Maximum A Posteriori Estimation for Multivariate Gaussian Mixture Observations of Markov Chains*. IEEE Transactions on Speech and Audio Processing, **Vol. 2**, Issue 2, pp. 291-297, April 1994.
12. GALLIANO S., GEUFFROIS E., BONASTRE J.-F., GRAVIER G., MOSTEFA D., CHOUKRI K., *Corpus Description of The ESTER Evaluation Campaign for The Rich Transcription of French Broadcast News.* Language Resources and Evaluation Conference, LREC 2006 Proceedings, 2006.