# Towards phonetically-driven hidden Markov models: Can we incoporate phonetic landmarks in HMM-based ASR?

*Guillaume Gravier, Daniel Moraru*

Équipe Metiss
Irisa, Rennes, France.
`http://www.irisa.fr/metiss`

## Abstract

Automatic speech recognition mainly relies on hidden Markov models (HMM) which make little use of phonetic knowledge. As an alternative, landmark based recognizers rely mainly on precise phonetic knowledge and exploit distinctive features. We propose a theoretical framework to combine both approaches by introducing phonetic knowledge in a non stationary HMM decoder. To demonstrate the potential of the method, we investigate how broad phonetic landmarks could be used to improve a HMM decoder by focusing the best path search. We show that, assuming error free landmark detection, every broad phonetic class brings a small improvement. The use of all the classes reduces the error rate from 22% to 14% on a broadcast news transcription task. We also experimentally validate that landmarks boundaries does not need to be detected precisely and that the algorithm is robust to non detection errors.

## 1. Introduction

In hidden Markov models (HMM) based speech recognition systems, the decoding process consists in compiling a graph which includes all the available sources of knowledge (language model, pronunciations, acoustic models) before finding out the best path in the graph in order to obtain the best word sequence

$$\hat{w} = \arg\max_{w} p(y|w)p(w) \ . \tag{1}$$

At the acoustic level, this approach relies on data-driven methods that learn from examples. Therefore, integrating explicit phonetic knowledge in such systems is difficult.

Alternately, various studies aimed at explicitly relying on phonetic knowledge to represent the speech signal for automatic speech recognition [1, 2, 3]. These approaches are most of the time based on the extraction of a set phonetic features, a.k.a. landmarks, on top of which a model, either rule based or statistical based, is build for the purpose of recognition. Phonetically-driven ASR relies on fine grain phonetic features such as onset and offset times [2] and distinctive features [1, 3]. However, in practice, automatically detecting such features might be difficult and error prone, in particular in the case of noisy signals or spontaneous speech.

This work is a preliminary study which aims at bridging these two paradigms in order to make use of explicit phonetic knowledge in the framework of HMMs. While landmark-based systems use phonetic landmarks as a feature describing the signal, the idea of our approach is to use landmarks in order to guide the search for the best path during Viterbi decoding in an HMM-based system. Hence, prior knowledge on the nature of the signal is used as *anchor points* during decoding. We will use indistinctly the two terms landmark and anchor to designate constraints on the search.

The aim of this study is twofold. The first aim is to define a theoretical framework to incorporate phonetic knowledge in HMM based systems using anchor points and to experimentally validate this approach. This framework allows for uncertainty in the landmark detection step, though this is not validated in the study as of now. The second aim is to study which landmarks effectively complements the data-driven knowledge embedded in HMM systems. We believe that detecting fine grain phonetic features is a particularly challenging problem – in spite of recent promising results on the detection of distinctive features, see *e.g.* [4, 5, 3] – while detecting broad phonetic features can be achieved with reasonably good performance [6, 7, 8]. Hence, to avoid problems related to the detection of fine grain features, we investigate if, and to what extent, broad phonetic landmarks can help.

in this paper, we first extend the Viterbi algorithm in order to incorporate prior knowledge carried out by landmarks. We then study the impact of broad phonetic landmarks in an ideal setting where landmarks are manually detected, with an emphasis on the temporal precision of the landmarks. Finally, we discuss some upcoming experiments whose results are expected to be presented at the workshop.

## 2. Landmark-driven Viterbi decoding

Most HMM-based systems rely on the Viterbi algorithm in order to solve Eq. (1), along with pruning techniques to keep the search tractable for large vocabularies. We briefly recall the basics of the Viterbi algorithm before extending this algorithm for the integration of phonetic anchors.

### 2.1. Beam-search Viterbi decoding

The Viterbi algorithm aims at finding out the best alignment path in a graph using dynamic programming (DP) on a trellis. The DP algorithm proceeds incrementally by searching for the best hypothesis reaching the state $(j, t)$ of the trellis according to

$$S(j,t) = \max_{i} S(i,t-1) + \ln(a_{ij}) + \ln(p(y_t|j)) \ , \tag{2}$$

where $j$ is the state in the decoding graph and $t$ the frame index in the observation sequence. In Eq. (2), $\ln(a_{ij})$ denotes the weight for the transition from state $i$ to $j$ in the graph[1] while

[1] Note that this weight actually combines the language model and the acoustic model probabilities for cross-word transitions.

$p(y_t|j)$ denotes the likelihood of the feature vector $y_t$ conditional to state $j$. Hence, $S(i,t)$ represents the score of the best partial path ending in state $i$ at time $t$.

In practice, not all paths are explored in order to keep the algorithm tractable on large decoding graphs. Unlikely partial hypotheses are pruned according to the score of the best path ending at time $t$.

### 2.2. Introducing anchors

Anchors can be considered as hints on what the best path is. For example, if a landmark indicates that a portion of an utterance corresponds to a vowel, then we can constrain the best path to be consistent with this piece of information since nodes in the decoding graph are linked to phonemes. One easy way to do this is to penalize, or even prune, all the paths of the trellis which are inconsistent with the knowledge brought by the vowel landmark. Assuming confidence measures are associated with the landmarks, the penalty should be proportional to the confidence.

Formally, the above principle can be expressed using non-stationary graphs, *i.e.* graphs whose transition probabilities are dependent on time. The idea is that if a transition leading to state $(i,t)$ of the trellis is inconsistent with the landmark knowledge, then the transition cost increases. In order to do this, we replace in (2) the transition weights $ln(a_{ij})$ by

$$\ln(a_{ij}(t)) = \ln(a_{ij}) - \lambda(t) I_j(t) \ . \tag{3}$$

$I_j(t)$ is an indicator function whose value is 0 if node $j$ is compatible with the available anchor information and 1 otherwise. The penalization term $\lambda(t) > 0$ reflects the confidence in the anchor available at time $t$, if any. Hence, if no anchor is available or if a node is consistent with the anchor, then no penalization is applied. In the opposite case, we apply a penalty where the higher the confidence in the landmark, the higher the penalty. In the extreme case where landmark detection is perfect, setting $\lambda(t) = \infty$, enables to actually prune paths inconsistent with the landmarks.

In (3), one can notice that the penalty term only depends on the target state $j$ and hence the proposed scheme is equivalent to modifying the state-conditional probability $p(y_t|j)$ to include a penalty. However, introducing the penalty at the transition level might be useful in the future to introduce phonological constraints or word-level constraints.

A by product of the proposed method is that decoding should be much faster with landmarks as adding a penalty will most likely result in inconsistent paths being pruned.

In this preliminary study, we use manually detected landmarks in order to investigate whether or not broad phonetic landmarks can help and to what extent in an ideal case. We will therefore set $\lambda(t) = \infty, \forall t$ in all the experiments described in section 4.

## 3. Baseline system

Before describing the experiments, we briefly present the data and baseline system used.

### 3.1. Corpus

Experiments are carried out on a radio broadcast news corpus in the French language. The data used is a 4 hour subset of the development data for the ESTER broadcast news rich transcription evaluation campaign [9]. The corpus mostly contains high-fidelity planned speech from professional radio speakers. Interviews, however, contain more spontaneous speech from non professional speakers, sometimes in degraded acoustic conditions.

The entire data set was labeled phonetically based on the reference orthographic transcription, using our ASR system to select pronunciation variants.

### 3.2. ASR system

Two reference systems were used in this study. Both systems are two-pass systems where a first pass aims at generating a word graph which is then rescored in a second pass with more sophisticated acoustic models. The two systems differ in the complexity of the acoustic models used for the word graph generation, the first system using context-independent models while word-internal context-dependent ones are used for the second system. Clearly, using landmarks to guide the decoding is more interesting when generating the word graph as it should enable better and smaller word graphs which already take into account the landmark knowledge. Therefore, the reason for comparing two systems for word graph generation is to determine to what extent phone models capture broad phonetic information.

Both transcription passes are carried out with a trigram language model. Monophone acoustic models have 114 states with 128 Gaussians per state while the word-internal triphone models have 4,019 distinct states with 32 Gaussians each. Cross-word triphones models are used for word graph rescoring, with about 6,000 distinct states and 32 Gaussians per state.

## 4. Broad phonetic landmarks

The experiments described in this section are performed using manually detected broad phonetic landmarks, the goal being to measure the best expected gain from the use of such landmarks. The main motivation for using this type of landmarks, as opposed to distinctive features, is that we believe that reliable and robust automatic broad phonetic landmark detectors can be build. For example, in [6, 7, 8] (to cite a few), good results are reported on the detection of nasals and vowels. Fricatives also seems relatively easy to detect using energy and zero crossing rate information. Moreover, we observed that the heap of active hypotheses in our ASR system most of time contains hypotheses corresponding to different broad phonetic classes. Though this is normal since hypotheses correspond to complete partial paths rather than to local decisions, this observation indicates that a better selection of the active hypotheses based on (locally detected) landmarks is bound to improve the results.

### 4.1. Landmark generation

Five broad phonetic classes are considered in this study, namely vowels, fricatives, plosives, nasal consonants and glides. Landmarks are generated from the available phonetic alignments obtained from the orthographic transcription. For each phone, a landmark corresponding to the broad phonetic class to which the phone belongs is generated, centered on the phone segment. The landmark duration is proportional to the phone segment length. In the first set of experiments, the landmark length is set to 50% of the phone segment length. We study in section 4.3 the impact of the landmark duration.

## 4.2. Which landmarks?

The first question to answer is what is the optimal improvement that can be obtained using each broad phonetic class separately. Results are given in table 1 for the monophone and triphone systems after the first and second pass, with each landmark type taken separately. Results using all the landmarks or a combination of vowel, plosive and fricative landmarks are also reported.

Results show a small improvement for each type of landmarks, thus clearly indicating that the transcription system is not misled by phones from a particular broad phonetic class. The best improvement is obtained with landmarks for glides, that correspond to highly transitory phones which are difficult to model, in particular because of co-articulation effects. More surprisingly, vowel landmarks yield a small but significant improvement, in spite of the fact that the phone models used in the ASR system do little confusions between vowels and other phones. This result is due to the fact that the DP maximization not only depends on the local state-conditional probabilities but also on the score of the entire path resulting in an hypothesis. In other words, even if the local probabilities $p(y_t|i)$ are much better for states corresponding to a vowel than for states corresponding to some other class, some paths, incompatible with the knowledge of a vowel landmark, might get a good cumulated score and are therefore kept in the heap of active hypotheses. Using the landmark-driven version of the Viterbi algorithm actually remove such paths from the search space, thus explaining the gain obtained with vowel landmarks.

Clearly, using all the available landmarks strongly improves the WER for both systems, the improvement being unsurprisingly better for the monophone-based system. One interesting point to note is that, when using all the landmarks, the two systems exhibit comparable levels of performance, with a slight advantage for the monophone system. This advantage is due to the fact that the word graph generated with the monophone system contains more sentence hypotheses than the one generated with the triphone system, though both graphs have roughly the same density. A last point worth noting is the rather good performance obtained after the first pass using the monophone system. This result suggest that combining landmark-driven decoding with fairly simple acoustic models can provide good transcriptions with a limited amount of computation. Indeed, the average number of active hypotheses, and hence the decoding time, is divided by a factor of four when using landmarks.

In a practical setting, the reliable detection and segmentation of a signal into broad phonetic classes is somewhat unrealistic, the detection of nasals and glides being a rather difficult problem. However, detecting vowels, plosives and fricatives seems feasible with a great accuracy. We therefore report results using only landmarks from those three classes (VPF results in table 1). Using such landmarks, a nice performance gain can still be expected, in particular with a monophone-based word graph generation.

These results show the optimal gain that can be obtained using broad phonetic landmarks as anchors in a Viterbi decoding, thus justifying further work on landmark detection.

## 4.3. Landmark precision

Two questions arise regarding the precision of the landmark detection step. The first question is to determine whether a precise detection of the landmark boundaries is necessary or not. The second question concerns the robustness to detection errors of the proposed algorithm.

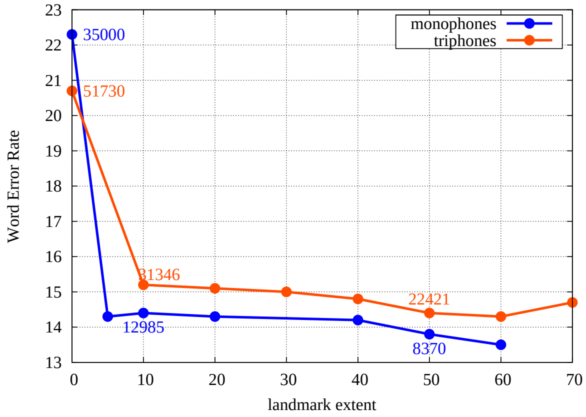

Figure 1: *WER (in %) as a function of the landmarks length, using all landmarks. The landmark length is defined as a fraction of the length of the phone which generated the landmark. Figures reported on the graph correspond to the average size of the active hypotheses heap during word graph generation.*

### 4.3.1. *Temporal precision*

Figure 1 shows the word error rate for the two systems as a function of the landmark extent, where the extent is defined as the relative duration with respect to the phone used to generate the landmark. An extent of 10 therefore means that the duration of a landmark is 0.1 times that of the corresponding phone. All the landmarks are considered in these experiments. Unsurprisingly, the longer the landmarks, the better the transcription. It was also observed that longer landmarks reduce the search space and yield smaller, yet better, word graphs. In spite of this, most of the improvement comes from the fact that landmarks are introduced, no matter their extent. Indeed, with a landmark extent of only 5%, the word error rate decreases from 22.3% to 14.3% with the monophone system. When increasing the landmark extent to 50%, the gain is marginal, with a word error rate of 13.9%. Note that with an extent of 5%, the total duration of landmarks corresponds to 4.4% of the total duration of the signal, and therefore landmark-based pruning of the hypotheses heap happens only for 4.4% of the frames. Similar conclusions were obtained using only the vowel landmarks. This is a particularly interesting result as it demonstrates that landmark boundaries do not need to be detected precisely. Reliably detecting landmarks on some very short portion of the signal (one or two frames) is sufficient to drive a Viterbi decoder with those landmarks.

### 4.3.2. *Robustness to non-detection errors*

In the absence of confidence measures, landmark-driven Viterbi is highly sensitive to detection errors. Clearly, false alarms, *i.e.* insertion and confusion errors, have detrimental effects on the system. However, miss detection errors are less disastrous. Therefore, automatic broad phonetic landmark detection systems should be designed to have as low as possible a false alarm rate. However, lower false alarm rates unfortunately imply higher miss detection rates. We tested the robustness of our landmark-driven decoder by simulating miss detection errors at various rates, assuming a uniform distribution of the errors across the five broad phonetic classes. Results show that the word error rate is a linear function of the miss detection rate.

Table 1: *Word error rate (in %) after each pass for the monophone and word-internal triphone systems, as a function of the landmarks used. The landmark ratio indicates the amount of signal (in %) for which a landmark is available.*

| landmarks | | none | all | VPF | vow. | plo. | fri. | nas. | gli. |
|---|---|---|---|---|---|---|---|---|---|
| landmark ratio | | | 43.6 | 34.6 | 18.3 | 9.0 | 7.3 | 2.8 | 6.2 |
| monophones | passe 1 | 29.2 | 15.3 | 21.7 | 26.6 | 26.5 | 27.5 | 27.8 | 25.1 |
| | passe 2 | 22.3 | 13.9 | 17.6 | 21.2 | 20.7 | 21.0 | 21.5 | 20.1 |
| triphones | passe 1 | 27.3 | 19.6 | 23.9 | 27.0 | 26.3 | 26.0 | 26.4 | 24.9 |
| | passe 2 | 21.3 | 15.0 | 18.2 | 20.7 | 20.4 | 20.3 | 20.7 | 19.6 |

For example, with the monophone system, the word error rate is 17.9% (resp. 15.8%) for a miss detection error rate of 50% (resp. 25%).

## 5. Discussion

The preliminary experiments reported in this summary are encouraging and prove that integrating broad phonetic landmarks in a HMM-based system can drastically improve the performance, assuming landmarks can be detected reliably. These results also validate the proposed paradigm for the integration of various sources of knowledge: phonetic knowledge via landmarks and data-driven knowledge acquired by the HMMs. However, results are reported in an ideal laboratory setting where landmark detection is perfect. The first step is therefore to work on robust detectors of broad phonetic landmarks, at least for vowels, plosives and fricatives, in order to validate the proposed paradigm in practical conditions.

A naive method for broad phonetic landmark detection was tested, based on broad phonetic class HMMs along with a trigram language model. For each of the five broad phonetic class, a context-independent left-right model with 3 states and 32 Gaussians per states was estimated on the training data of the ESTER corpus. These models were then used for broad phonetic segmentation with a trigram language model, resulting in an accuracy of 76.6[2]. Assuming landmarks extracted from this broad phonetic segmentation with a landmark extent of 20% of the segment length, the amount of landmark time that is not correctly labeled is 10.8%, which did not prove sufficiently low to help the ASR system. Indeed, we used the landmarks associated with sufficiently long segments as anchors to prune inconsistent hypotheses in our system, assuming long segments are more reliable than short ones. On a 1 hour show, the baseline word error rate of 22.3% increased to 23.2% with vowel, plosive and fricative landmarks and 24.3% considering all the landmarks. However, the segmentation system is naive in the sense that it relies on the same features and techniques than the ASR system and therefore does not bring any new information. It seems clear that a better broad phonetic segmentation system, based on different features, can be devised. Moreover, *segmentation* may not be the best strategy for landmark detection and techniques that differs from the HMM framework (*e.g.* MLP, SVM) should be used for the *detection* of broad phonetic landmarks. Results with more robust landmark detectors and using confidence measures will be presented at the workshop.

Finally, let us conclude this discussion with two remarks. First, we believe that mixing the landmark paradigm with data-driven methods offers a great potential to tackle the problem of robustness. In this sense, broad phonetic landmarks seems a reasonable choice to achieve robustness. In particular, we think that human perception might actually follow a similar scheme as the one presented here, where landmarks are used to disambiguate sounds and words. Second, we would like to stress that the framework defined in this paper for the integration of phonetic knowledge in a HMM based system is not limited to speech recognition with landmarks. The framework offers a way to integrate knowledge in a DP algorithm in a general way and has many application fields such as multimodal fusion or audiovisual speech recognition.

## 6. Acknowledgments

This work was partially financed in the framework of the Action Concertée Incitative Masses de Données Demi-Ton – Multimodal Description for the Automatic Structuring of TV Streams.

## 7. References

[1] S. A. Liu, "Landmark detection for distinctive feature-based speech recognition," Ph.D. dissertation, Massachusetts Institute of Technology, 1995.

[2] A. Juneja, "Speech recognition based on phonetic features and acoustic landmarks," Ph.D. dissertation, University of Maryland, 2004.

[3] *Landmark-based speech recognition: report of the 2004 John Hopkins Summer Workshop*. John Hopkins University, Center for Language and Speech Processing, 2005.

[4] E. McDermott and T. Hazen, "Minimum classification error training of landmark models for real-time continuous speech recognition," in *Proc. IEEE Intl. Conf. Acoust., Speech, Signal Processing*, vol. 1, 2004, pp. 937 – 940.

[5] J. G. K. Schutte, "Robust detection of sonorant landmarks," in *Eurospean Conf. on Speech Communication and Technology – Interspeech*, 2005, pp. 1005 – 1008.

[6] M. Chen, "Nasal landmark detection," in *Intl. Conf. Speech and Language Processing*, 2000, pp. 636–639.

[7] A. Howitt, "Vowel landmark detection," in *Intl. Conf. Speech and Language Processing*, 2000.

[8] J. Li and C.-H. Lee, "On designing and evaluating speech event detectors," in *European Conference on Speech Communication and Technology – Interspeech*, 2006, pp. 3365–3368.

[9] S. Galliano, E. Geoffrois, J.-F. Bonastre, G. Gravier, D. Mostefa, and K. Choukri, "Corpus description of the ESTER evaluation campaign for the rich transcription of french broadcast news," in *Language Resources and Evaluation Conference*, 2006.

[2] Most of the errors are due to the glides while vowels are well detected. Surprisingly, fricatives are not very well detected.