Using Broad Phonetic Classes to Guide Search in Automatic Speech Recognition

Stefan Ziegler, Bogdan Ludusan, Guillaume Gravier

Using Broad Phonetic Classes to Guide Search in Automatic Speech Recognition

Stefan Ziegler, Bogdan Ludusan, Guillaume Gravier

CNRS – IRISA, Campus de Beaulieu, 35042 Rennes, France

{stefan.ziegler, bogdan.ludusan, guillaume.gravier}@irisa.fr

Abstract

This work presents a novel framework to guide the Viterbi decoding process of a hidden Markov model based speech recognition system by means of broad phonetic classes. In a first step, decision trees are employed, along with frame and segment based attributes, in order to detect broad phonetic classes in the speech signal. Then, the detected phonetic classes are used to reinforce paths in the search process, either at every frame or at phonetically significant landmarks. Results obtained on French broadcast news data show a relative improvement in word error rate of about 2% with respect to the baseline.

Index Terms: Viterbi decoding, broad phonetic classes, landmarks

1. Introduction

State of the art hidden Markov model (HMM) based automatic speech recognition (ASR) often suffers a significant drop in performance when being confronted with unknown variations in speech, like new speaking styles or accents. Since the articulatory aspects of human speech production are well understood nowadays and HMM-based ASR captures only few of this phonetic knowledge in its models, several studies have been aiming at exploiting phonetic knowledge to achieve more robust ASR.

Some approaches to speech recognition, like Lexical Access from Features (LAFF) [1] and event-based system recognition [2], rely exclusively on phonetic knowledge by detecting perceptual important events (landmarks) in the speech signal, events which provide information about articulatory gestures and the basic sound contrasts in speech. Phonetic knowledge can also serve as additional source of knowledge inside probabilistic frameworks, like in [3], where landmark and segment-based information is combined inside a probabilistic ASR system. The integration of articulatory information into a decoding graph was also studied in [4], where broad phonetic classes (BPCs) were used to concentrate computational efforts during decoding inside reliable regions of speech. In [5], support vector machines are used to detect phonetic landmarks for performing lattice rescoring in a two-pass ASR system.

In this paper, we extend the framework initially proposed in [6], which used landmarks in the form of BPCs to guide a Viterbi decoding process, in order to obtain improved word graphs. In that study, manually detected landmarks served as anchor points during decoding to strictly prune paths incompatible with the presented landmark information. Our first contribution lies in presenting a BPC classification system that organizes selected attributes into decision trees to detect BPCs in the speech signal. We compare two approaches to BPC classification. The first one estimates the BPC for every frame in the signal, while the second approach predicts the BPC only for certain landmark-frames. These landmarks are obtained by estimating the most salient points of articulatory gestures. Furthermore, we replace the existing phonetically driven decoding by a more flexible implementation that enhances paths according to broad phonetic knowledge.

The paper is organized as follows: In the first part, we present the phonetically driven Viterbi decoding and provide details about our BPC classification system. The evaluation section presents the experimental setup and reports recognition results, while the paper concludes with an outlook on future work.

This work was partially supported by the Agence Nationale de la Recherche in the framework of the ASH project. The authors would also like to thank Régine André-Obrecht and Jérôme Farinas for providing them with the implementation of the forward-backward divergence method.

2. ASR Driven by Phonetic Knowledge

In HMM-based ASR, the Viterbi algorithm uses dynamic programming to keep track of the best path reaching state (j, t) by computing a score

$$S(j,t) = \max_i S(i,t-1) + log(a_{ij}) + log(p(y_t|j)) + R(j,t). \quad (1)$$

$S(i, t-1)$ is the score for being in state i at time $t-1$, $log(a_{ij})$ denotes the transition log-probability from state i to j and $log(p(y_t|j))$ represents the likelihood of the feature vector y_t conditional to state j. $R(j,t)$ introduces knowledge about the correct path at frame t into the Viterbi decoding by reinforcing state j by $R(j,t) = \lambda_{j,t} \cdot R_{max}$. $\lambda_{j,t}$ is a binary value, indicating whether state j is compatible with the external knowledge source and R_{max} acts as the enhancement factor, limiting the influence of $R(j,t)$ on the overall score. If there is no knowledge about the correct path at frame t, $R(j,t) = 0$ for all j. Broad phonetic knowledge can easily be introduced into the decoder, since every state j is directly linked to one BPC.

BPCs organize phonemes into different classes according to similarities in their articulatory gestures. The phonemes of one BPC share the same acoustic properties, while the classes are maximal discriminant towards each other. Our intention is to add a BPC classifier parallel to the existing HMM-based ASR system, that estimates the presence of BPCs in the signal. For each frame for which a BPC is detected, all states that are member of this BPC will be activated with $\lambda_{j,t} = 1$, while the remaining states are set to $\lambda_{j,t} = 0$. In the following section, we will provide details about our BPC classification system.

3. Detection of Broad Phonetic Information

Our BPC classifier estimates the probability $p_{BPC_c}(t)$ of $c = 8$ classes: vowels, nasals, approximants, plosives, fricatives (further divided by voiced and unvoiced) and a final group repre-

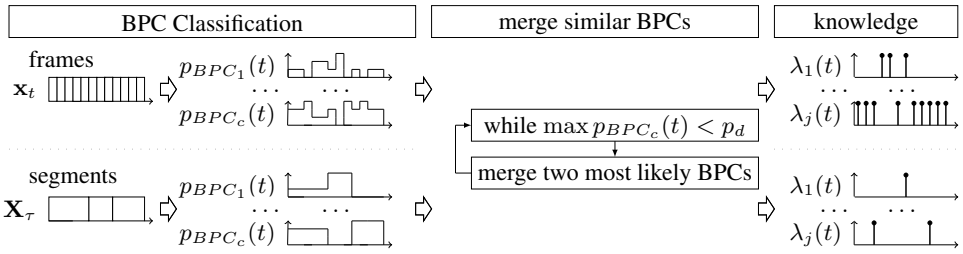

Figure 1: Frame and segment-based BPC detection. Both approaches estimate BPC probabilities for each frame, respectively segment. For each segment, the frame for which the detected segment-BPC has the highest frame probability is chosen as the landmark.

senting the non-speech events inbreath and silence. Introducing these automatically detected BPCs into the decoding is not straightforward, since there is a trade-off between providing as much phonetic information as possible, while preventing the insertion of wrong information. We study two factors that influence this trade-off.

The first factor is the place where $p_{BPC_c}(t)$ is estimated: either for every frame in the speech signal or only for selected landmark frames, where the acoustic correlates of articulatory gestures are assumed to be most salient. While the frame-based approach has the possibility to provide the most amount of phonetic information, unreliable parts of speech, for example phoneme transitions, are likely to misguide decoding due to misclassified BPCs. In contrast to that, landmarks do not predict the BPC for every frame, but could still guide towards the correct path by correctly enhancing at least one frame inside each phoneme.

The second factor concerns the confusion between acoustically similar BPCs, like approximants and vowels. To reduce the phonetic errors introduced into the decoding, we set a minimum BPC probability p_d and merge the two most likely BPCs into a more general phonetic class, until the probability of the most likely BPC exceeds p_d. By increasing p_d, the phonetic information at every classified frame might become very broad, while the errors introduced into the decoding will decrease.

In section 3.1, we will focus on the extraction of attributes for the frame and landmark-based BPC classifier. While both approaches use frame-based acoustic attributes for prediction, landmarks rely additionally on segmentation to estimate articulatory movements. The details about the mapping of those attributes onto their respective BPCs by using decision trees are provided in section 3.2.

3.1. Frame and Segment-based Attributes

At the frame level, the acoustic content of the BPCs is modeled by m acoustic parameters (APs) $x_{t,m}$ that are extracted for each frame t. Short-term contextual information is provided by adding a small context window of one frame around t, resulting in $m \cdot 3$ attributes per frame.

To extract landmarks from the speech signal, we divide the speech signal into a sequence of segments, using the forward-backward divergence method [7], which sequentially detects segment boundaries by comparing the parameters of a long term with that of a short term auto-regressive (AR) model. A significant change in the AR-models is supposed to indicate a change in the articulatory movement. Segmentation groups the frame-based representation of speech $\mathbf{x} = \mathbf{x}_1, \ldots, \mathbf{x}_t$, into a sequence of segments $\mathbf{X} = \mathbf{X}_1, \ldots, \mathbf{X}_\tau$ with $\mathbf{X}_\tau$ containing all frames $\mathbf{x}_t$ inside the τ-th segment. The acoustic content of a segment is described by the 3 frames at the center of each segment, to capture the spectral transitions of the articulatory movement. Contextual information of subsequent segments is again provided by adding the attributes of the two neighbouring segments to a total of $9 \cdot m$ attributes for each segment. After the classifier has estimated the BPC probabilities of a segment, one frame inside the segment is chosen as the landmark, by selecting the frame where the most probable segment-BPC has the highest frame probability.

3.2. Decision Trees for BPC Prediction

The design of our BPC classification system is motivated by landmark-based approaches to speech recognition, like LAFF [1], which provide a set of expert rules that map acoustic cues to their respective phonetic classes. By employing decision trees, we aim at automatically organizing the attributes into a similar set of binary rules. Decision trees provide reliable probability estimates, since the class probabilities are derived from the distribution of the classes at the end nodes of the tree and can accommodate for nominal information, in our case information about the channel bandwidth. We compensate potentially unstable behaviour of single trees by bootstrap aggregating.

While Mel-frequency cepstral coefficients (MFCCs) are state of the art for the acoustic modeling of phonemes, BPCs might equally be discriminated by broader, yet more robust APs, like low-level features describing general spectral shapes. We employ correlation-based feature selection [8] to obtain the most discriminative APs from a large amount of attributes, which include MFCCs (with first and second order derivatives), spectral shapes (e.g. spectral centroid), temporal statistics (e.g. zero-crossing rate), energy measurements (e.g. bark bands) and formant information.

The speaker diarization information of the speech recognizer provides the possibility to normalize the APs on a common range for each speaker. Our normalization method first divides the attributes of the m-th AP for each speaker into q quantiles. Normalization is performed by replacing every attribute $x_{t,m}$ with $x_{t,m}^{norm} \in \{1, 2, \ldots, q\}$ by $x_{t,m}^{norm} = \arg\min_q |x_{t,m} - c_{q,m}|$. $c_{q,m}$ is the median of the values inside the q-th quantile of the m-th attribute for the current speaker. Constant effects of the channel and speaker on $x_{t,m}$, often modeled as changes in mean and variance, will not affect the quantiles of the distribution. Besides increasing the robustness, this method reduces memory requirements, since $x_{t,m}^{norm}$ requires only 7 bits of memory, assuming $q = 128$. Furthermore, in combination with decision trees, only 128 different split positions for each node have to be considered during training, which equally reduces computation time.

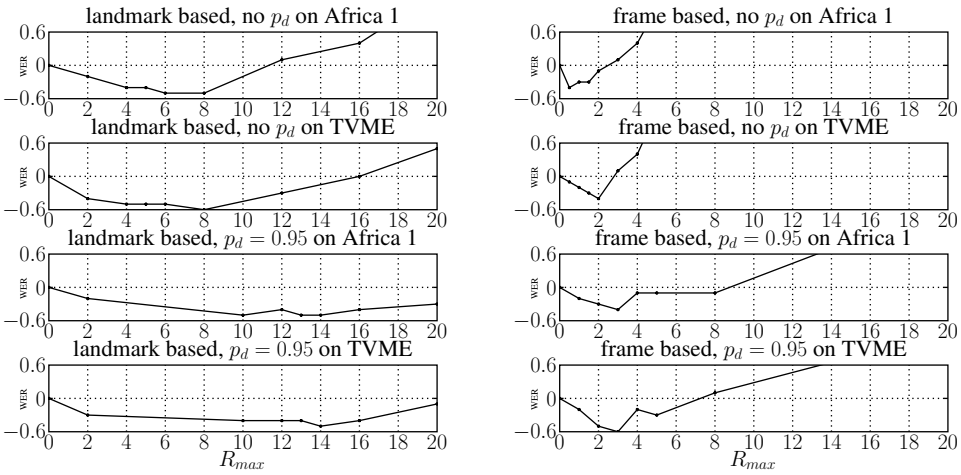

Figure 2: WER improvement of two news shows (135 min Africa 1 and 60 min TVME) as a function of R_{max}.

4. Experiments

The speech recognizer used in this paper is a two-pass system, with the first pass generating a word graph and the second pass rescoring the previously obtained graph using more complex models. The acoustic models employed in the first pass are word-internal triphones with 4,019 distinct states and 32 Gaussians per state and word trigrams are used as language model. The rescoring pass uses 4-grams as language model and cross-word triphone models with 6,000 states and 32 Gaussians each. The acoustic models are trained on 150 hours of mainly planned speech in studio broadcast environments. As proposed in [6], we employ phonetic constraints only in the first pass, to generate better word graphs. The word error rates (WERs) reported are calculated on the final hypotheses obtained after all passes.

Decision trees are built using the WEKA toolbox [9] with information gain as splitting criterion, reduced-error pruning and 30 bootstrap aggregating iterations on a 50% sample of the training data. The APs are extracted using YAAFE [10] and the Snack Sound toolkit [11] and we additionally add the nominal attribute *bandwidth*, obtained from the diarization system, to discriminate telephone from wide band speech. Trees are trained on 1.67 million instances from 26 hours training data, for both frame and segment-based training. For the frame-based classifier, each training instance corresponds to the $3 \cdot m$ attributes from the three frames at the center of each segment, while each instance of the segment-based classifier additionally contains the $6 \cdot m$ attributes of the neighbouring segments. A segment was only used for training, if the large majority (e.g. 70%) of the segment was inside a phonemic boundary.

Experiments are conducted on radio broadcast news in French language from the ESTER2 campaign [12], which contain news shows with regular broadcast speech (RFI), but also difficult tasks like debates (Inter) or accentuated speech (TVME and Africa 1). Our development set contains 16 recordings from the ESTER2 development set, covering three different news shows (135 min Africa 1, 60 min Inter, 60 min TVME) and the test set contains 23 recordings (90 min Africa 1, 70 min RFI, 60 min Inter, 60 min TVME) from the ESTER2 test set.

normalization	m APs	accuracy
not normalized	35	69.7
mean variance	35	71.7
quantiles ($q = 128$)	35	71.8
quantiles ($q = 128$)	62	72.3

Table 1: Comparison of classification performance on the development set.

4.1. BPC Classification

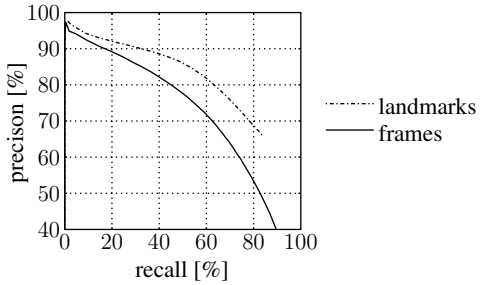

Figure 3: Precision-recall curve, evaluated on the phoneme level for the frame and landmark-based BPCs.

Feature selection resulted in 62 attributes, containing only few MFCCs, mostly derivatives. For computational reasons, we first used the 35 best ranked APs with 10 bagging iterations and 30% sample size to optimize classification parameters like pruning strategy or splitting criterion, considering only the wide band speech during training and testing. Table 1 compares the classification performance of the best settings obtained, but with different normalization methods of the APs. The best performing quantile normalization method is also displayed using the full 62 APs.

In Figure 3, we compare the precision-recall curve of the frame-based BPCs to the landmark BPCs. For both cases, the frames having a probability estimate $p_{BPC_c}(t)$ higher than the current threshold were used for evaluating the performance on a phoneme level. A phoneme was considered as correctly detected, as long as all classified frames inside the phoneme matched the phoneme label. As soon there was one falsely predicted frame inside the phoneme boundaries, it was regarded as a misclassified phoneme. If no frame inside a phoneme was above the current threshold, the phoneme was considered as a missed phoneme. While the segmentation generally over-segmented phonemes (see [7]), some parts of speech were

under-segmented, especially strongly co-articulated phonemes, which led to a maximum recall for landmarks of 84%. It can be seen that using our proposed method for landmark detection increases the phoneme precision, compared to the frame-based BPCs at equal recall.

Table 2 illustrates the relation between phonetic information, as the the average cardinality of the detected BPCs introduced into the system, and phoneme errors for several values for p_d. Increasing p_d reduces classification errors at the cost of less phonetic information, since many BPCs were merged into very general phonetic classes.

	AVG BPC size		incorrect phonemes [%]	
p_d	landmark	frame	landmark	frame
-	8	8	31	85
0.8	14	16	13	45
0.95	22	24	4	17

Table 2: Relation between average size of the BPCs detected and phoneme errors on the development set for different p_d. The total number of speech and non-speech symbols is 40.

4.2. Speech Recognition

	baseline		frame		landmark	
p_d	-		0.8		0.6	
R_{max}	-		2		8	
broadcast	dev	test	dev	test	dev	test
Inter	21.9	18.7	21.7	18.4	21.6	18.6
RFI	-	17.6	-	17.3	-	17.2
Africa 1	45.1	31.5	44.6	30.7	44.7	30.8
TVME	30.1	24.1	29.6	24.2	29.6	24.4

Table 3: WER[%] for the optimal p_d and R_{max} on the development and test set.

Figure 2 displays the improvement in WER as a function of R_{max} for four different experimental settings. Each setting was applied on two radio stations of the development set. The left column corresponds to recognition driven by landmark BPCs, while the recognition with frame-based BPCs is on the right. p_d was not employed in the upper two rows, while $p_d = 0.95$ was used in the lower two rows. If no p_d option is used, frame-based BPCs are very sensitive towards the choice of R_{max} and improvement in WER is only obtained inside a small window. Furthermore, the two different radio shows differ in their optimal R_{max}. As soon as less errors are introduced into the system, by using $p_d = 0.95$ or introducing BPCs only at landmarks, the exact choice of R_{max} gets less important, since R_{max} rarely enhances the wrong path. For all values of p_d being tested on the development set, both, frame and landmark-driven decoding improved the WER at the optimal R_{max}.

The pairing R_{max} and p_d was optimized on the development set for both, the landmark and frame-based approach and the two optimal pairs were then applied on the test set. The overall improvement in WER with respect to the 23.5% WER of the baseline of the test set was 0.4% for the frame-based BPCs and 0.3% for the landmark-based BPCs. The statistical significance of the WER improvement was tested using a Wilcoxon signed rank test and it proved to be significant at the 5% level, for both, the landmark and frame-based case. Radio Inter achieved the least gain on the development set with 0.2% (frame) and 0.3% (landmarks) but confirmed this small gain on the test set. RFI was not included in the development set, but gained 0.3% and 0.4%. Africa 1 performed well and improved the WER 0.8% and 0.7% on the test set. While TVME gained 0.5% for both approaches on the development set, this was not confirmed on the test set, where two recordings containing partly non-native speakers increased the WER.

5. Conclusions

In this paper, we proposed a method for driving the Viterbi decoding of a HMM-based ASR system by automatically detected BPCs. We compared a frame-based and a landmark-based approach to BPC detection using decision trees and we limited phonetic errors by merging similar BPCs according to estimated BPC probabilities. While both approaches achieved a similar improvement on broadcast news shows, landmarks proved to be more robust towards the choice of the maximum enhancement factor.

With the promising results phonetically driven decoding was able to achieve in this study, one part of our future work is to improve the existing BPC classification. Furthermore, efforts will be made towards combining frame and landmark-based BPC classification and exploring the integration of additional knowledge sources. Finally, we plan to explore methods that introduce knowledge about BPCs into the word graph of the rescoring step.

6. References

[1] K. N. Stevens, "Toward a model for lexical access based on acoustic landmarks and distinctive features," *J. Acoust. Soc. Am.*, vol. 111, pp. 1872–1891, 2002.

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

[3] J. R. Glass, "A probabilistic framework for segment-based speech recognition," *Comput. Speech Lang.*, vol. 17, pp. 137–152, 2003.

[4] T. N. Sainath, "Island-driven search using broad phonetic classes," in *Proc. of ASRU-2009*, 2009, pp. 287–292.

[5] M. Hasegawa-Johnson, J. Baker, S. Borys, K. Chen, E. Coogan, S. Greenberg, A. Juneja, K. Kirchhoff, K. Livescu, S. Mohan, J. Muller, K. Sonmez, and T. Wang, "Landmark-based speech recognition: report of the 2004 Johns Hopkins summer workshop," in *Proc. of ICASSP'05*, 2005, pp. 213–216.

[6] G. Gravier and D. Moraru, "Towards phonetically-driven hidden Markov models: can we incorporate phonetic landmarks in HMM-based ASR?" in *Proc. of NOLISP'07*, 2007, pp. 161–168.

[7] R. André-Obrecht, "A new statistical approach for the automatic segmentation of continuous speech signals," *IEEE Trans. on ASSP*, vol. 36, pp. 29–40, 1988.

[8] M. Hall, "Correlation-based feature selection for machine learning," Ph.D. dissertation, The University of Waikato, 1999.

[9] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA data mining software: an update," *ACM SIGKDD Explorations Newsletter*, vol. 11, pp. 10–18, 2009.

[10] B. Mathieu, S. Essid, T. Fillon, J. Prado, and G. Richard, "YAAFE, an easy to use and efficient audio feature extraction software," in *Proc. of ISMIR-2010*, 2010.

[11] K. Sjlander, "The snack sound toolkit," www.speech.kth.se/snack, 2004, accessed on 12 dec 2011.

[12] S. Galliano, G. Gravier, and L. Chaubard, "The ESTER 2 evaluation campaign for the rich transcription of French broadcasts," in *Proc. of INTERSPEECH-2009*, 2009, pp. 1149–1152.