Towards a new speech event detection approach for landmark-based speech recognition

Stefan Ziegler, Bogdan Ludusan, Guillaume Gravier

▶ **To cite this version:**

Stefan Ziegler, Bogdan Ludusan, Guillaume Gravier. Towards a new speech event detection approach for landmark-based speech recognition. SLT - Workshop on Spoken Language Technology, 2012, United States. hal-00758424

HAL Id: hal-00758424
https://hal.science/hal-00758424

Submitted on 28 Nov 2012

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

TOWARDS A NEW SPEECH EVENT DETECTION APPROACH FOR LANDMARK-BASED SPEECH RECOGNITION

Stefan Ziegler, Bogdan Ludusan, Guillaume Gravier

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

ABSTRACT

In this work, we present a new approach for the classification and detection of speech units for the use in landmark or event-based speech recognition systems. We use segmentation to model any time-variable speech unit by a fixed-dimensional observation vector, in order to train a committee of boosted decision stumps on labeled training data. Given an unknown speech signal, the presence of a desired speech unit is estimated by searching for each time frame the corresponding segment, that provides the maximum classification score. This approach improves the accuracy of a phoneme classification task by 1.7%, compared to classification using HMMs. Applying this approach to the detection of broad phonetic landmarks inside a landmark-driven HMM-based speech recognizer significantly improves speech recognition.

Index Terms— speech event detection, landmark-driven ASR

1. INTRODUCTION

In state-of-the-art hidden Markov model-based (HMM) automatic speech recognition (ASR), speech is modeled as a sequence of phone-segments, often referred to as the beads-on-a-string model of speech [1]. In contrast to that, acoustic-phonetic or event-based approaches to speech recognition model speech as a stream of asynchronous phonetic events, which have to be further processed to obtain a higher level speech representation [2, 3, 4, 5]. Many of these approaches require the detection of phonetic events as time instances, referred to as landmarks.

There are numerous studies concerning the detection of phonetic events in the speech signal, which can be divided into two general groups. The first group uses expert knowledge to derive detection rules from various signal representations (e.g., [2]). The second group uses a classification and detection approach, where labeled data is converted into the desired phonetic feature representation and classifiers are trained to map acoustic observations onto phonetic speech units (e.g., [3, 4]). Employing these classifiers sequentially on the speech signal results in detection functions indicating the presence of the desired speech units. These classifiers operate usually on a frame basis, extracting a fixed-dimensional observation vector for each frame, eventually considering a small context window. On the one hand, this fixed observation space enables the use of non-linear and non-parametric classifiers like support vector machines [3, 4] or multilayer perceptrons for classification, which possess good discrimination abilities and can be trained efficiently. On the other hand, predicting speech units by observing a fixed-size fraction of the speech signal does only partly model time-variable speech units, including phonemes and many articulatory feature units.

We are interested in the problem of extracting a fixed-dimensional observation vector from time-variable speech units, to discriminate them using a classifier that requires a fixed number of observations. To obtain this observation vector, we use a maximum-likelihood segmentation method, to force each labeled speech unit into three spectrally homogeneous parts from which the observation vector is extracted. In our work, this observation vector is used to train speech units with boosted decision trees and to predict unknown segments during testing. To obtain a detection function, indicating the locations of a speech event by its local maxima, we search for each frame the segment that maximizes the classification score at this frame. This method is applied to a recently proposed landmark-driven ASR framework, that uses phonetic landmarks to guide the search in an HMM-based ASR system [6].

The paper is organized as follows. First, we present the details of our proposed method and review boosting with an ensemble of weak learners, as well as landmark-driven ASR. Evaluation is carried out on a phoneme classification task and on landmark-driven speech recognition experiments and we conclude with an outlook on future work.

2. PROPOSED SYSTEM FOR SPEECH EVENT DETECTION

In event-based speech recognition, phonetic speech units are often trained and classified at the frame level. The classification scores of successively predicted frames correspond to a

This work was partially supported by the Agence National de la Recherche in the framework of the ASH project. Furthermore, we would like to thank Christian Raymond for advices concerning the use of his boosting software.

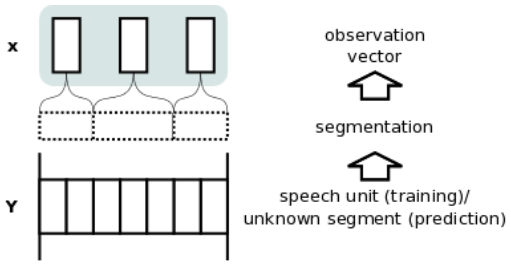

Fig. 1. Diagram of the proposed method of using segmentation to extract a fixed-dimensional observation vector from a labeled speech unit (training) or an unknown segment (testing).

detection function, that indicates the presence of the speech event, usually by its magnitude. Landmarks, corresponding to single time instances marking the most salient points of speech events, are then obtained by post-processing the detection profiles, varying from simple peak-picking to processing with statistical models (e.g. [3]). To capture spectral transitions and temporal information, usually a concatenation of parametrized speech frames create a fixed-dimensional observation space for each frame for classification. But since speech units are commonly modeled as time-variable segments, a fixed number of frames will either capture only fractions of the spectral information of the speech unit or contain misleading information about previous or following speech events. If the detection function is obtained by inaccurate modeling of the desired speech event, it is difficult to recover the loss in information by post-processing noisy and erroneous detection profiles.

In the following, we present a method that overcomes this shortcomings by extracting a fixed-dimensional observation vector $\mathbf{x}$ from a given time-variable sequence of parametrized speech $\mathbf{Y} = \mathbf{y}_1, \ldots, \mathbf{y}_t$ (section 2.1), corresponding to any desired speech unit. We are able to keep a fixed-dimensional observation space, since we reduce every speech segment to three subsegments, similar to a three-state HMM. This observation vector enables the use of any desired classification method, while generalizing all time-variable speech units. In our case, we apply a committee of boosted decision stumps to train a classier $F(\mathbf{x})$ (section 2.2). $F(\mathbf{x})$ is then used to generate a detection function $\mathbf{d}_k = d_{k,1}, \ldots, d_{k,t}$ for k desired speech units, with the profile of $\mathbf{d}_k$ indicating the presence of this speech unit in the unknown signal. For each k-th detection function, a set of l_k time instances $L_k = \{t_1, \ldots, t_{l_k}\}$ is extracted, representing the exact locations of the speech event associated with speech unit k (section 2.3).

2.1. Maximum-likelihood segmentation applied to speech units

In the following, we apply the maximum-likelihood speech segmentation approach described in [7] to the segmentation of speech units. Given the parametrized frame-based representation of a labeled speech unit $\mathbf{Y} = \mathbf{y}_1, \ldots, \mathbf{y}_n$, for example a sequence of Mel-frequency cepstral coefficients (MFCCs) vectors, where n corresponds to the length in frames of the speech unit, we aim at finding the segment borders b_2 and b_3 which segment the unit into $i = 3$ segments $(\mathbf{y}_{b_1}, \ldots, \mathbf{y}_{b_2-1})$, $(\mathbf{y}_{b_2}, \ldots, \mathbf{y}_{b_3-1})$ and $(\mathbf{y}_{b_3}, \ldots, \mathbf{y}_n)$, with $b_1 = 1$.

The optimal segment borders are considered to be the borders minimizing the intra-segment distortion of each segment. As proposed in [7], we measure the intra-segment distortion as the accumulation of distances from each frame inside the segment to the segment-centroid μ_i. Using the euclidean distance between frames as a distance measure, the task is to find the two frames b_2 and b_3 given $b_1 = 1$ and $b_4 = n + 1$ according to

$$b_2, b_3 = \underset{b_2,b_3}{\arg\min} \sum_{i=1}^{3} \sum_{n=b_i}^{b_{i+1}-1} \|\mathbf{y}_n - \mu_i\| \,;\; b_1 = 1, b_4 = n+1. \tag{1}$$

This corresponds to a shortest-path problem, which can be solved for each segment using dynamic programming.

The observation vector $\mathbf{x}$ for each speech unit is obtained by first concatenating the centroids of the obtained segments $\left[\mu_1^T, \mu_2^T, \mu_3^T\right]$. Additional attributes added to $\mathbf{x}$ are the length of each segment, to capture temporal information, and the three intra-segment distortion measures $\sum_{n=b_i}^{b_{i+1}-1} \|\mathbf{y}_n - \mu_i\|$ to measure the homogeneity of each segment.

2.2. Boosting for speech unit classification

Having extracted a fixed-dimensional observation vector for each variable-length speech unit, speech unit models can be trained using the desired classifier. In this work, we boost decision stumps with the multiclass AdaBoost.MH algorithm [8].

Boosting is an ensemble learning method that iteratively learns weak learners $h_m(\mathbf{x})$ to classify instances, that were for the most part misclassified by the previous learner. The error of learner $h_m(\mathbf{x})$ determines its weight α_m. The outputs of m weak learners $h(\mathbf{x})$ are combined to a strong classifier $H(\mathbf{x})$, providing a score for every class k with $H(\mathbf{x}) = \sum_m \alpha_m h_m(\mathbf{x})$.

A set of weak learners provides a very flexible framework during prediction, since the trade-off between computation time and precision can be adapted by changing the number of weak learners. Furthermore, the use of decision trees allows to insert nominal attributes in the future, like information about gender or bandwidth.

The training data used in this work includes over 200 hours of speech, or about 8.5 million labeled speech units. Training on the entire data set is usually not feasible, since most algorithms need the data to fit into memory and training time increases non-linearly. We follow the approach proposed in [9], which is to divide our training set into J partitions and

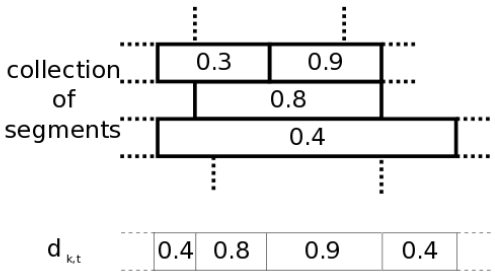

Fig. 2. Simplified example illustrating the extraction of a detection function $\mathbf{d}_k$ from a given set of segments. Each segment is displayed with its prediction score. The maximum scores at each frame t are mapped onto $d_{k,t}$.

to train a classifier $H_j(\mathbf{x})$ on each partition. The final classification consists in averaging the output scores of the full committee of classifiers:

$$F(\mathbf{x}) = \frac{\sum_{j=1}^{J} H_j(\mathbf{x})}{J}. \quad (2)$$

2.3. From classification to detection

As stated in the introduction of section 2, $F(\mathbf{x})$ is used to obtain a detection function $\mathbf{d}_k = d_{k,1}, \ldots, d_{k,t}$. If $F(\mathbf{x})$ is trained to predict single frames $\mathbf{y}_t$ of parametrized speech, $d_{k,t}$ directly corresponds to the prediction score of class k at frame t. Since our classifier provides segment and not frame-based prediction scores, we provide a new method to generate detection profiles from a collection of predicted segments.

Intuitively, in order to obtain $d_{k,t}$, one has to search for the segment containing t, that maximizes the likelihood of the given speech unit at frame t. Given a collection of overlapping segments, with each segment corresponding to a hypothetical speech unit, we can directly compare the score of variable-length segments, using our proposed observation-vector. A collection of segments does not correspond to a graph, so that single segments might not have a connection to preceding or subsequent segments (see simplified example in Figure 2). With this work being a preliminary study, we do not consider the task of finding a collection of suitable prior segments, which is desirable to reduce the computational costs for generating $d_{k,t}$. Instead, we use an exhaustive collection of segments by predicting the observations extracted from all possible segments inside the speech signal, up to a maximum segment length of 300ms. Using fast classification like boosting, and an effective implementation for the solution of equation 1, this stays computationally feasible. The value of the detection function $d_{k,t}$ at t is calculated as follows:

$$d_{k,t} = \max F_k(\mathbf{x}_{(s,e)});\ s \leq t \leq e, e - s \geq i - 1. \quad (3)$$

$F_k(\mathbf{x})$ is the score of the k-th detection function given the observation $\mathbf{x}$. s and e correspond to the first and last frame of a segment and $\mathbf{x}_{(s,e)}$ to the observation vector extracted from this segment. The minimum number of frames in a segment is $i = 3$.

The obtained detection function indicates the positions of the k-th speech event in the speech signal by its local maxima. Since we put most effort into the accurate modeling of the speech units, we expect clear indications of the speech events, without requiring sophisticated post-processing. The following simple peak picking algorithm can be applied to a system of discrete speech units, like phonemes or broad phonetic classes (BPCs). Given the detection functions for k speech units, we extract landmarks in three steps:

1. We first collect the time instances corresponding to local maxima for each detection function $\mathbf{d}_k$ and associate each local maximum t with its magnitude $d_{k,t}$. Since a local maximum corresponds to a segment of at least three frames, the central frame t of a segment is always selected as the time instance t of the local maximum.

2. If several units k share the same local maximum t, only the landmark with the highest magnitude $d_{k,t}$ is kept at t.

3. Some of the remaining landmarks might correspond to a local maximum at a very low magnitude. Therefore, at each landmark, all units k with a score $d_{k,t}$ bigger than the landmark magnitude are also activated at that time instance. Afterwards, each set $L_k = \{t_1, \ldots, t_{l_k}\}$ contains the final collection of l_k discrete time instances signaling the presence of the k-th speech event.

If a local maximum t of speech unit k already corresponds to the highest score $d_{k,t}$ at that time frame, step 3 will have no effect. If not, this step will merge several speech units into broader units, effectively signaling the possible presence of several units at the same time.

3. LANDMARK-DRIVEN HMM-BASED ASR

We use the presented speech event detection approach to extract broad phonetic landmarks for a recently proposed landmark-driven HMM-based ASR framework [6]. In this framework, broad phonetic landmarks are used to guide the Viterbi decoding of the first pass of a HMM-based speech recognizer, which we briefly summarize in the following.

The Viterbi algorithm searches the best path reaching state (j, t) by computing a score

$$S(j,t) = \max_i S(i, t-1) + log(a_{ij}) + log(p((y)_t|j)) + R(j,t), \quad (4)$$

where $log(a_{ij})$ is the transition probability and $log(p(y_t|j))$ corresponds to the acoustic likelihood. $R(j,t)$ enhances paths through states j that are compatible with phonetic landmarks by $R(j,t) = \lambda_{j,t} \cdot R_{max}$. With states corresponding

to triphones, checking the compatibility is trivial, since each state can be directly linked to one BPC. $\lambda_{j,t}$ is a binary indicator, that becomes 1 if state j is compatible with the landmark at time t and 0 if there is no landmark at all. R_{max} limits the influence of landmarks onto the overall score and has to be determined experimentally.

4. EXPERIMENTS

Our experiments focus on two objectives. First, we are interested in evaluating whether our proposed approach can capture the acoustic information of speech units and compare the classification performance to existing approaches. Second, the proposed approach will be used to extract broad phonetic landmarks to guide the search of an HMM-ASR system as described in section 3.

4.1. Experimental setup

The speech corpus used in the experiments corresponds to radio broadcast news in the French language from the ESTER2 broadcast transcription evaluation campaign [10]. Our training set consists of over 200 hours of speech, the development set of 3 hours and the test set of 4.5 hours from 4 broadcast shows (radio Africa1, radio Inter, radio RFI, radio TVME). The test set is divided into $J = 12$ non-overlapping partitions for a committee of 12 classifiers. Classification experiments are run on the development set, while speech recognition is carried out on the test set. Speech labels are obtained by forced-alignment.

The speech recognizer used for recognition experiments 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. We use the tool bonzaiboost[1] for the training of the ensemble of weak classifiers .

4.2. Phoneme classification

While landmark detection is primarily used for the detection of phonetic events, our proposed method is designed to learn any time-variable speech unit. This includes phonetically motivated units, as well as phonemes. We decided to do classification experiments on a phoneme classification task, since this is a challenging task, considering the fine acoustic distinctions between phonemes. Our French phoneme alphabet consists of 40 phonemes, including 5 non-speech events.

HMM-monophone models serve as the baseline phoneme classifier, because of their time warping ability, which makes them perfectly suitable for discriminating time varying speech units. The models employed correspond to monophone 3-state left-to-right HMMs with 64 diagonal-covariance Gaussian components per state. HMMs are trained on the full training set using the same speech parametrization as the boosted ensembles, which are 39-dimensional MFCC vectors, composed of 13 MFCC coefficients with first and second order derivations. To compare the abilities of the classifiers to capture the acoustic properties of each phoneme, we evaluate on a pure classification task, i.e. all observation vectors were extracted using the known phoneme borders obtained by forced alignment, either for training or predicting the associated phoneme. HMM-classification was performed by force-aligning each model to the known phoneme borders and comparing the obtained likelihoods.

First, we trained four classifiers on only one partition (i.e. 1/12-th) of the training data for comparison, before creating the full committee of 12 classifiers. The first classifier uses the three concatenated frames at the temporal center of each phoneme as observation vector $\mathbf{x}$ for training and testing (see Figure 4). Training and predicting only the center of each phoneme is supposed to minimize errors due to coarticulation effects. We compare this approach to our proposed method, which we run using weak learners of different depth (classi-

[1] developed by Christian Raymond, available at http://bonzaiboost.gforge.inria.fr

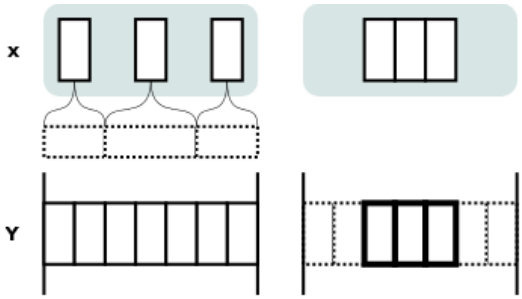

Fig. 4. Proposed observation vector for the phoneme classification experiments. On the left, the observation vector extracted for each phoneme as proposed in this paper (classifier 2, 3, 4 and 6). On the right, the observation vector as a concatenation of subsequent speech frames at the center of each phoneme (classifier 1).

trained on $1/12$ of training set

#	classifier	accuracy
1	boosting, concatenated frames (depth=2)	55.7
2	boosting, proposed (depth=1)	57.8
3	boosting, proposed (depth=2)	63.0
4	boosting, proposed (depth=3)	62.8

trained on full training set

#	classifier	accuracy
5	HMM (64 gaussians)	65.9
6	committee, proposed ($J = 12$, depth=2)	67.6

Table 1. Classifiers of the phoneme classification task. The classifier is described by its classification method (e.g. boosting) and observation vector (e.g. proposed).

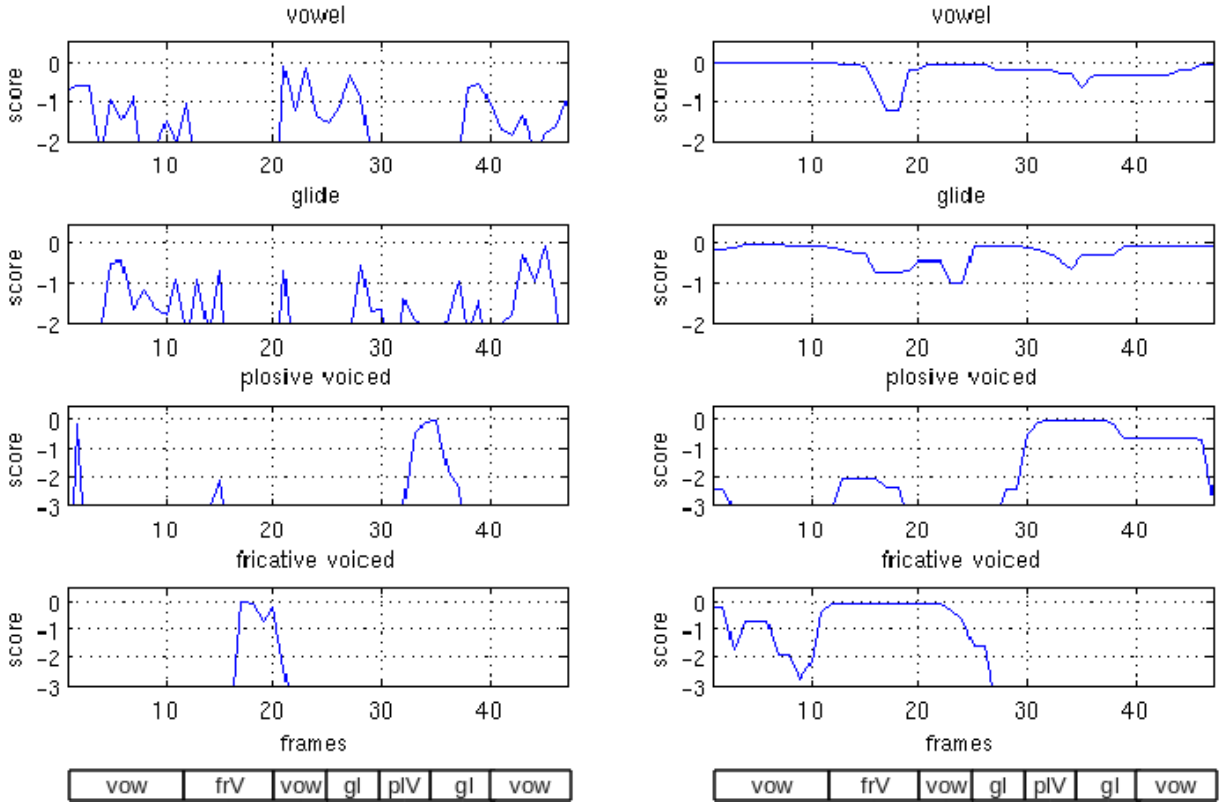

Fig. 3. Detection profiles of the four broad phonetic units vowels, glides, plosives (voiced) and fricatives (voiced) for the french word *aujourd'hui* (uttered during a broadcast news show) and its corresponding broad phonetic annotation. On the left, every frame of the detection function was predicted using classifier 1 (see Table 1). On the right, the proposed method was used to generate the detection score at each frame (using classifier 3). While vowels, plosive and fricative are well indicated for both classifiers, the detection function for classifier 1 is more noisy, especially for the vowels. Glides seem to be more clearly represented on the right side, which can be due to the fact, that glides correspond to slow articulatory movements, which can be difficult to capture by concatenated frames.

fier 2, 3 and 4). For all experiments the number of boosting rounds was limited to 3000. Table 1 displays the classification accuracy on the development set as the percentage of correctly classified phonemes. Using the proposed observation vector for boosting increases the performance by 7.3% (using decision trees with depth=2) compared to training on concatenated frames. While boosting on one partition of the training set does not outperform the accuracy of HMMs, applying the committee of 12 boosted weak ensembles improves by 1.7% compared to HMMs.

	classifier		
	1	3	6
number of landmarks	1,352k	574k	608k
phoneme errors [%]	67.7	20.2	18.0
missed phonemes [%]	4.6	12.0	8.8
AVG BPC size [phonemes]	13	20	20

Table 2. Statistics of extracted landmarks on the whole development set for three classifiers. AVG BPC size corresponds to the average number of phonemes provided at each landmark.

4.3. Landmark-driven ASR

To use our phoneme classifier for broad phonetic landmark detection, we simply derived a collection of detection functions $\mathbf{d}_j$ for $j = 7$ BPCs vowels, nasals, glides, fricatives (voiced and unvoiced) and plosives (voiced and unvoiced) by scoring every frame $d_{j,t}$ with the score $d_{k,t}$ of the phoneme k, that has the maximum score among all phonemes of this BPC at frame t.

We extracted landmarks as proposed in section 2.3 for classifiers 3 and 6 from Table 1. For classifier 1 we equally extracted landmarks, but using the detection function obtained by predicting each individual frame. Figure 3 compares de-

broadcast	WER baseline	WER landmark-driven
Inter	18.7	18.6
RFI	17.6	17.3
TVME	24.2	23.8
Africa1	31.5	30.8

Table 3. Speech recognition performance driven by broad phonetic landmarks. Landmarks are extracted using classifier 6.

tection profiles obtained by predicting single frames using classifier 1 and by employing our proposed method (classifier 3). Table 2 contains information about the landmark accuracy. A phoneme error corresponds to a phoneme with at least one landmark that misclassifies this phoneme. As one can expect, directly extracting local maxima from frame based detection functions leads to more than twice as much landmarks compared to the proposed method, because of the noisy detection profile (see also Figure 3), resulting in many incorrect landmarks. It should be noted, that this is partly due to our landmark extraction algorithm, which is designed to avoid any post processing, like smoothing or heuristic peak-picking algorithms. The high average number of active phonemes at each landmark is due to many landmarks that consist of several merged BPCs.

For the landmark-driven ASR experiments according to section 3, we used the landmarks obtained with classifier 6. First, the development set was used to tune R_{max} (see equation 4). The optimal R_{max} was then employed for the landmark-driven decoding on the test set. The results in Table 3 show an improvement for all 4 broadcast shows tested, compared to the baseline which does not include landmarks. The improvement in word-error-rate (WER) varies from 0.1 (radio Inter) to 0.7 (radio Africa1). The overall WER of the test set was 23.1%, compared to a 23.5% baseline, and a Wilcoxon signed rank showed it to be significant at the 5% level. Compared to the WER obtained by the phonetically guided decoding presented in [6], where BPCs were trained on selected frames and decoding was guided by predicting the BPC of every frame, no broadcast show of the test set was degraded by the use of broad phonetic information.

5. CONCLUSIONS

In this work, we presented a new method for the detection of variable-length speech units. We used segmentation to obtain a fixed-dimensional observation vector for each speech unit to train a committee of boosted decision stumps for speech unit classification. To detect speech units in an unknown signal, we searched for each time frame the corresponding segment, that provides the maximum classification score for the desired speech unit. This approach improved the accuracy of a phoneme classification task compared to HMM-phoneme classification, as well as the WER of a hybrid HMM-based landmark-driven ASR framework, compared to its HMM-based baseline.

With the promising results we were able to obtain by employing our proposed framework, there are several directions for future research. First, the proposed fixed-dimensional observation space could be refined, by considering different segmentation and attribute-extraction strategies. Second, the step from classification to detection by evaluating all possible segments should be replaced by an efficient search for a suitable collection of prior segments. While boosting decision stumps performed well on the presented classification task, applying other classification techniques might further improve speech unit classification and detection.

6. REFERENCES

[1] M. Ostendorf, "Moving beyond the beads-on-a-string model of speech," in *Proc. of ASRU-1999*, 1999, pp. 79–84.

[2] 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.

[3] A. Jansen and P. Niyogi, "Point process models for event-based speech recognition," *Speech Communication*, vol. 51, no. 12, pp. 1155–1168, 2009.

[4] A. Juneja, *Speech recognition based on phonetic features and acoustic landmarks*, Ph.D. thesis, University of Maryland, College Park, 2004.

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

[6] S. Ziegler, B. Ludusan, and G. Gravier, "Using broad phonetic classes to guide search in automatic speech recognition," in *Proc. of INTERSPEECH-2012*, 2012.

[7] A.K.V. SaiJayram, V. Ramasubramanian, and T.V. Sreenivas, "Robust parameters for automatic segmentation of speech," in *Proc. of ICASSP'02*, 2002, p. 513.

[8] Y. Freund, R. Schapire, and N. Abe, "A short introduction to boosting," *Japanese Society For Artificial Intelligence*, vol. 14, no. 5, pp. 771–780, 1999.

[9] N.V. Chawla, T.E. Moore, L.O. Hall, K.W. Bowyer, P. Kegelmeyer, and C. Springer, "Distributed learning with bagging-like performance," *Pattern Recognition Letters*, vol. 24, no. 1, pp. 455–471, 2003.

[10] 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.