# UNSUPERVISED ACOUSTIC SUB-WORD UNIT DETECTION FOR QUERY-BY-EXAMPLE SPOKEN TERM DETECTION

*Marijn Huijbregts, Mitchell McLaren and David van Leeuwen*

Centre for Language and Speech Technology, Radboud University Nijmegen, The Netherlands

{marijn.huijbregts, m.mclaren, d.vanleeuwen}@let.ru.nl

## ABSTRACT

In this paper we present a method for automatically generating acoustic sub-word units that can substitute conventional phone models in a query-by-example spoken term detection system. We generate the sub-word units with a modified version of our speaker diarization system. Given a speech recording, the original diarization system generates a set of speaker models in an unsupervised manner without the need for training or development data. Modifying the diarization system to process the speech of a single speaker and decreasing the minimum segment duration constraint allows us to detect speaker-dependent sub-word units. For the task of query-by-example spoken term detection, we show that the proposed system performs well on both broadcast and non-broadcast recordings, unlike a conventional phone-based system trained solely on broadcast data. A mean average precision of 0.28 and 0.38 was obtained for experiments on broadcast news and on a set of war veteran interviews, respectively.

***Index Terms***— Spoken term detection, zero resource speech recognition, acoustic sub-word unit generation, speaker diarization

## 1. INTRODUCTION

Modern large vocabulary continuous speech recognition (LVCSR) systems require large collections of orthographically transcribed speech data for training their statistical language and acoustic models. It is not always feasible to construct such data sets due to the time and expense associated with the annotation of large quantities of audio. This is typically the case for 'low resource' languages for which performing automatic speech recognition (ASR) is economically less profitable. For example, due to the lack of automatic transcription tools, only a few percent of the recordings of 4,000+ endangered languages, currently being made by linguists, can be analyzed [1].

Even if sufficient training data is available for a particular language to develop standard broadcast news speech recognition systems, high quality ASR for other speech styles of that language cannot be guaranteed. For example, sufficient resources are available to perform LVCSR on Dutch broadcast news data [2], but we are still struggling to recognize interview material consisting of unprepared speech in various acoustic conditions. Recently we have developed a system for automatic transcription of a collection of interviews with Dutch war veterans. Despite the manual annotation of small portions of audio for acoustic model adaptation and training special in-domain language models, the performance of this system left considerable room for improvement.

Given that limited annotated training data is available for many LVCSR tasks, new techniques are needed that are robust to such low training resources. Ideally, an ASR system would require no training data at all. For Speech Activity Detection (SAD) and speaker diarization we have shown that such an approach is actually possible [3, 4]. In these cases we use a Hidden Markov Model (HMM) based segmentation and clustering set-up, in which the Gaussian Mixture Models (GMM) needed for each class are trained and iteratively refined on the test data itself. In order to do the same within a standard ASR framework, we would need to be able to automatically generate acoustic models, a dictionary and a language model. It is the first step, automatic generation of acoustic models, that presents the focus in this work. We extend our speaker diarization system in our attempt to accomplish this.

Speaker diarization is the task of segmenting and clustering speech based on speaker identity. Our diarization system is able to do this without the need of external training data. It clusters the test audio based on the most salient acoustic variation within the recording — the variation between speakers. If we only process speech from one single speaker and relax a few system constraints, we expect the system to model the acoustic variation at a smaller granularity, e.g., the phone or sub-word level. We hypothesize that the acoustic sub-word units generated in this case may be an appropriate substitution for conventional acoustic models.

In this paper we investigate the automatic generation of speaker-dependent acoustic models on the test audio itself using a system originally developed for speaker diarization. We will evaluate the models in a spoken term detection task. Because we do not use resources such as a lexicon and language model to perform large vocabulary continuous speech recognition, we will perform query-by-example spoken term detection. With this type of spoken term detection, an example fragment of a recording is used as query to find other, similar fragments in the recording. Although the applicability of such a search-by-example system may appear limited, we find a useful application in the veteran interview collection where single speakers are talking for several hours in a single recording and for which traditional LVCSR did not provide satisfactory search facilities.

The remainder of this paper is organized as follows. Section 2 describes studies that are related to this work. Our proposed approach is detailed in section 3. Section 4 describes the experimental set-up followed by a discussion in section 5 on how we plan to expand our system towards speaker- and recording-independent acoustic sub-word unit modeling.

## 2. RELATED WORK

A number recent studies have been carried out on automatically obtaining acoustic sub-word units. We will briefly discuss them in this section.

The methodology of discovering words or sub-words from the

raw speech signal is an important issue in studies of language acquisition in babies. In a recent study, an algorithm for unsupervised word discovery was presented that automatically generated phone-like sub-word units [5]. The segmentation and clustering approach employed in this study was based on dynamic time warping comparisons. In [6], a GMM was trained on speech from multiple speakers and the posteriors of each Gaussian in the mixture were used as units in a dynamic time warping framework. In [7], unsupervised learning of acoustic sub-word units was achieved using divisive clustering. In this study, an extended method for allophone learning was used to go from one cluster (all speech) towards sufficient multiple acoustic sub-word units for Japanese.

The work in the three aforementioned studies is similar to our work. All papers were successful in automatically finding a set of acoustic sub-word units, each using different clustering methods — dynamic time warping, training a GMM to be used in a posteriorgram and applying an extended form of the successive state splitting algorithm. In contrast to the first two clustering methods, the approach taken in this work is to re-align the data iteratively over the models and to refine the models with each iteration (see section 3). For speaker diarization this iterative approach has proven to be very effective. Although in the third method [7], the data is also iteratively re-aligned over the models, a divisive clustering algorithm is used instead of the agglomerative algorithm applied in this study. A potential advantage of agglomerative clustering over divisive clustering is that in general, considerably more iterations are applied using agglomerative clustering, allowing for better refinement of the models. Further, in our clustering approach, the number of components of each GMM scales with the amount of data being processed. This also works very well for speaker diarization.

We were not able to implement all three techniques and compare them on our evaluation set. Instead, we chose to implement solely the GMM-based approach for comparison due to its ability to find acoustic sub-word units across speakers [6]. This means that we can also use this method as a baseline for future research when we will try to link sub-words of speakers to generate a speaker independent set of units.

A clear description of a system set-up for query-by-example spoken term detection can be found in [8]. In this study the output of a conventional phone recognizer is transformed into a *posteriorgram*, a matrix with the posterior probabilities of all phones for all time frames of the recording. Such a posteriorgram is made for both the query and the recording. Next, for each time frame in the query, the distance to all frames in the recording is calculated. Dynamic time warping is then used to find relevant 'documents' for each query. We will use the framework of posteriorgram calculation and dynamic time warping for our own retrieval set-up using our automatically generated sub-word units instead of the phone posteriors used in [8].

## 3. AGGLOMERATIVE CLUSTERING APPROACH

Our query-by-example spoken term detection system consists of three steps. First, acoustic sub-word units are generated using the modified speaker diarization system. Next, using the models for the sub-word units, a posteriorgram is formed. Finally, for each query a template is constructed and, with the use of dynamic time warping, a list of relevant documents is generated. In this section we will first describe the steps of our approach and then we will describe three systems that we built to establish a baseline for comparison to our system. In the following section we will describe how we applied the systems to a retrieval experiment.

### 3.1. Unsupervised acoustic sub-word unit detection (UASUD)

Before we perform unsupervised acoustic sub-word unit detection, we first run our standard speech activity detection and speaker diarization system to filter out all non-speech and to obtain the speech segments labeled by speaker identity [3, 4]. We then perform unsupervised acoustic sub-word unit detection (UASUD) using a model-based agglomerative clustering approach in which speech data is first divided into a large number of clusters and these clusters are merged pairwise until the desired number of clusters is reached. The models that are needed during UASUD are not generated on an external training set but on the audio that is being processed itself.

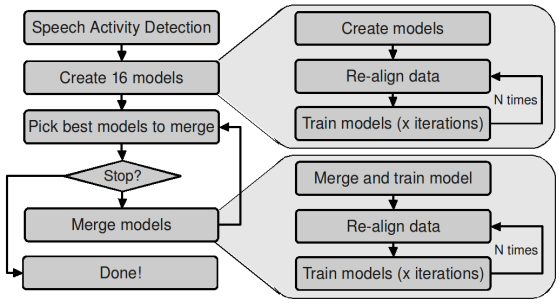


**Fig. 1**. *A schematic representation of the UASUD-algorithm. The steps for creating the initial models and merging the models each consist of a number of training and re-alignment iterations.*

Figure 1 presents the five steps of the algorithm. In the first step, speech activity detection and diarization, all non-speech audio is removed from the data and the speech segments are labeled on speaker identity so that the remaining steps can be performed on the speech segments of each individual speaker. In the second step, initial acoustic sub-word unit models are generated in an iterative manner. This involves random assignment of the speech data to each GMM, followed by iteratively training the models and re-aligning the data. In the third step, a distance metric is used to determine which two models are most similar which are subsequently merged in the fourth step. The third and fourth steps are repeated until the desired number of clusters is reached.

For Viterbi re-alignment the GMMs are organized in an HMM topology depicted in figure 2. Each acoustic sub-word unit is represented by a string of four states that share the single GMM as probability density function (pdf). This topology ensures that the minimum duration of each segment after the Viterbi search is 40 ms.

**Fig. 2**. *The HMM topology of the UASUD-system. Each sub-word unit is represented by a string of four states that uses a single GMM.*

In order to find the two acoustic units that are most similar, the Bayesian Information Criterion (BIC) distance metric is used [9]. This metric compares the two models $M_i$ and $M_j$ with a third model $M_{ij}$ that is trained on the combined data of $M_i$ and $M_j$. The advantage of this metric is that it does not require a threshold that needs to be tuned on a development set as long as the number of free parameters in $M_{ij}$ (the number of gaussians) equals the number of free parameters in $M_i + M_j$ [10, 11]. This metric was first used for acoustic change detection in [12].

### 3.2. Comparison to speaker diarization

The UASUD-system described above remarkably resembles the design of a speaker diarization system. In fact, our system is derived from our speaker diarization system [4]. There are, however, several distinct differences between the proposed UASUD-system and

our diarization system. First, the minimum duration imposed by the HMM topology is 4 states, compared to 250 for the diarization system. Second, in addition of the 19 MFCC features used in the diarization system, the UASUD-system uses the first and second order derivatives of the features, thereby allowing our system to capture more of the smaller granularity dynamics in speech. Thirdly, we initialize with 207 clusters (approx 3.5 times the final number of acoustic units) whereas we use 16 for speaker diarization, assuming application in scenarios with 2–10 speakers. Finally, the speaker diarization system has an additional stopping criterion due to the number of speakers in the recording being unknown. For our UASUD-system, we use a fixed number of 57 final clusters for the reasons detailed in section 3.4.

### 3.3. Search by dynamic time warping on posteriorgrams

In [8], a phonetic posteriorgram is defined as a time-vs-class matrix representing the posterior probability of each phonetic class for each time frame of a recording. Similar to the approach in [8], we create a posteriorgram for both the recording and each query. However, instead of using phone probabilities to create the posteriorgrams, we use the probabilities of the acoustic sub-word units.

From the query posteriorgram and the recording posteriorgram, we create a similarity matrix containing a similarity score for each vector $\mathbf{q}$ in the query posteriorgram with each vector $\mathbf{x}$ in the recording posteriorgram. The similarity score $S$ is determined as follows:

$$S(\mathbf{q}, \mathbf{x}) = \log(\mathbf{q} \cdot \mathbf{x}) \quad (1)$$

Using the similarity matrix we perform Dynamic Time Warping (DTW) to retrieve fragments in the recording that contain speech similar to the query. We constrain the DTW search to a local time warping ratio of maximum 2. That is, each frame in the query sequence is assigned to maximum two frames in the search sequence and vice versa.

### 3.4. Alternative query-by-example systems

We have developed three alternative systems for comparison to our proposed system. All systems make use of the same posteriorgram/DTW framework for search, but the posteriorgram probabilities are obtained in three different ways. For the first system, the *MFCC* system, the probabilities are substituted by the MFCC feature vectors of the modified diarization system. Each of the 57 MFCC components is normalized over the entire recording so that all values fall between zero and one.

The second system, the *GMM* system, is our implementation of the system from [6] (see section 2). We train the GMM up to 57 Gaussians so that the systems are comparable.

Finally, we have created a system using posterior probabilities of our Dutch phone models, as obtained from our LVCSR BN-system developed for Dutch [4].

## 4. EXPERIMENTAL SET-UP AND RESULTS

We test the proposed system and the three baseline systems on two sets of data. We use all 78 speech segments, totaling 7.5 minutes in duration, of the anchor-person in a Dutch broadcast news recording to see how the algorithms work on clean studio speech. We then test the systems on 15 veteran interviews. These interviews are tabletop microphone recordings of Dutch war veterans, who are senior citizens. Each interview is two hours long. From each interview we have manually annotated 2.5 minutes of speech for evaluation purposes. We performed a forced-alignment with our regular Dutch LVCSR BN decoder to obtain the exact start and end time of each word. Note that although we only evaluate on 2.5 minutes of each recording, the systems process each interview in its entirety.

Normally, for spoken document retrieval tasks, only content words are used as queries because these are the words of interest. As we are interested in the acoustic modeling capabilities, we determine the performance on *all* words in the annotated parts of the recordings, and do not limit ourselves to content words.

For a retrieval experiment we need the concept of a 'document' and for this we use the speech segments obtained from the SAD module. We use each annotated word as a query and request the system to output all other speech segments that contain this word. DTW is used to find the optimal relevance score for each document and we then sort the documents according to this score.

We evaluate the system performance in terms of the Mean Average Precision (MAP).

### 4.1. Results

Table 1 contains the results of the four systems on the broadcast news data and on the 15 interviews. It shows that the phone search worked well for in-domain broadcast news data, but not for the interview data. This was expected as the LVCSR system was developed for broadcast news data for which it obtained a Word Error Rate (WER) of 34.9% [4], whereas the system achieved a 63.5% WER on the interview data. The three other systems are not based on models trained on external data and, consequently, did not suffer from the same performance drop as the phone-based system. Searching directly on the MFCC features was suboptimal on both test sets. The straightforward GMM-based system performed surprisingly well, however our proposed UASUD-system performed consistently better on the two tasks. The improvement was 3.7% relative on broadcast news and 5.6% relative on the interview data.

**Table 1**. Results of the four systems.

| Experiment | MAP on BN | MAP on interviews |
|---|---|---|
| MFCC | 0.25 | 0.32 |
| phone | 0.27 | 0.06 |
| GMM | 0.27 | 0.36 |
| UASUD | 0.28 | 0.38 |

To analyze if the output of the UASUD-system and the GMM-based system are different, we have plotted the individual average precision scores of the two system outputs against each other (figure 3). Although the MAP of the two systems is similar, the cloud in figure 3 shows that the average precision is different for most queries. This suggests that the fusion of the two techniques, such as often is done in the field of speaker recognition, could lead to improved performance [13].

The MFCC-based and the phone-based systems are not affected by the duration of each recording as they do not train or tune any models on the data. However, this is not the case for the GMM-based system and the UASUD-system. On the one hand, long recordings provide more data to train on and hence leads to potentially richer models, but on the other hand, in longer recordings more variation in speech will be observed than is necessary for the short evaluation segments. This potentially makes it harder to generate a small set of models that can be successfully applied for spoken term detection. To check if the length of the recording influences the performance of the system, we repeated the interview experiment for the GMM-based system and the UASUD-system with the adjustment that only the 2.5 minutes of evaluation data needed to be processed. The mean

average precision for the UASUD-system was not affected by this (0.38). The MAP of the GMM-based system actually improved to 0.37 (0.36 on the original test). It is hard to predict whether this indicates that the GMM-based system is less robust to long recordings. Experiments on much larger audio collections are needed to prove this hypothesis.

**Fig. 3**. *The average precision of each query for the UASUD-system (horizontal) and the GMM-based system (vertical) on the veteran interview collection.*

## 5. DISCUSSION

In this paper we have shown that with only a few small adjustments, it is possible to use a standard speaker diarization system to automatically learn acoustic sub-word units that are suitable for use in a spoken term detection system. From the four systems that we evaluated in this study our approach performed best, offering a marginal but consistent performance improvement over the GMM-based method. It is remarkable that the divisive GMM approach and the agglomerative UASUD approach consistently give rise to quite similar MAP, where the per-keyword performance is quite uncorrelated.

This study was a first step in the direction of performing ASR with the use of as little supervised training resources as possible. We were able to automatically generate speaker-dependent acoustic sub-word units that proved useful in a spoken term detection framework. In future research we will extend our system to be able to generate speaker independent sub-word units. We will investigate whether it is possible to do this the same way as we do for diarization of large archives: perform clustering for each speaker and link the acoustic units between the different speakers using generalized detection techniques [14]. It will further be interesting to see if our assumption is true that the difference in performance between this approach and the GMM-based method grows for larger data sets.

Another step that we will need to take in future research is automatically generating pronunciation lexicons for our acoustic sub-word units. We plan to 'identify' words by searching for often appearing recurrent sequences of sub-word units in our data. The authors of [15] managed to do this for relatively long sequences using conventional phone models. It will be interesting to find out if we are able to perform similar experiments with our automatically generated acoustic sub-word units.

Orthographically labeling the generated 'words' will be the next challenge. We realize that the solution for this challenge is far away; the problem can only be solved once the other steps are successfully taken. Hopefully it will be possible to find word labels by comparing sequences of 'words' with the actual words in textual newspaper archives.

## 6. ACKNOWLEDGEMENTS

This paper is based on research carried out for the BATS project within the research program IM-Pact funded by IBBT and ICTRegie.

## 7. REFERENCES

[1] L. Boves, R. Carlson, E. Hinrichs, D. House, S. Krauwer, L. Lemnitzer, M. Vainio, and P. Wittenburg, "Resources for speech research: Present and future infrastructure needs," in *proceedings of Interspeech*, 2009.

[2] Judith Kessens and David van Leeuwen, "N-best: The northern- and southern-dutch benchmark evaluation of speech recognition technology," in *Interspeech*, Antwerp, Belgium, August 2007.

[3] Marijn Huijbregts and Franciska de Jong, "Robust speech/non-speech classification in heterogeneous multimedia content," *Speech Communication*, vol. In Press, Corrected Proof, 2010.

[4] Marijn Huijbregts, *Segmentation, diarization and speech transcription : surprise data unraveled*, Ph.D. thesis, Enschede, November 2008.

[5] L.F.M. ten Bosch and B. Cranen, "A computational model for unsupervised word discovery," in *proceedings of Interspeech*, Antwerp, Belgium, 2009.

[6] Yaodong Zhang and James Glass, "Towards multi-speaker unsupervised speech pattern discovery," in *proceedings of ICASSP*, Dallas, 2010.

[7] Balakrishnan Varadarajan, Sanjeev Khudanpur, and Emmanuel Dupoux, "Unsupervised learning of acoustic sub-word units," in *proceedings of HLT'08*, Columbus, Ohio, 2008.

[8] T.J. Hazen, W. Shen, and C. White, "Query-by-example spoken term detection using phonetic posteriorgram templates," in *proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding*, December 2009.

[9] G. Schwartz, "Estimating the dimension of a model," *The Annals of Statistics*, vol. 6, no. 2, pp. 461–464, 1978.

[10] Jitendra Ajmera, H. Bourlard, I. Lapidot, and I. McCowan, "Unknown-multiple speaker clustering using HMM," in *proceedings of the International Conference on Spoken Language Processing (ICSLP)*, Denver, Colorado, USA, 2002.

[11] Xavier Anguera, *Robust Speaker Diarization for Meetings*, Ph.D. thesis, Universitat Politecnica De Catalunya, 2006.

[12] Shaobing S. Chen and P. Gopalakrishnan, "Speaker, environment and channel change detection and clustering via the bayesian information criterion," in *proceedings of DARPA Broadcast News Transcription and Understanding Workshop*, Virginia, USA, 1998.

[13] Niko Brummer, Lukas Burget, and et al., "Fusion of heterogenous speaker recognition systems in the stbu submission for the nist speaker recognition evaluation 2006," in *proceedings of IEEE TASLP*, September 2007.

[14] Marijn Huijbregts and David van Leeuwen, "Towards automatic speaker retrieval for large multimedia archives," in *proceedings of 3d ACM Workshop on Automated Information Extraction in Media Productions*, Firenze, Italy, October 2010.

[15] Aren Jansen, Kenneth Church, and Hynek Hermansky, "Towards spoken term discovery at scale with zero resources," in *proceedings of Interspeech*, Makuhari, Japan, 2010.