Automatic discovery of subword units and pronunciations for automatic speech recognition using TIMIT

George Goussard and Thomas Niesler
Department of Electrical and Electronic Engineering
Stellenbosch University, South Africa
trn@sun.ac.za

***Abstract*—We address the automatic generation of acoustic subword units and an associated pronunciation dictionary for speech recognition. The speech audio is first segmented into phoneme-like units by detecting points at which the spectral characteristics of the signal change abruptly. These audio segments are subsequently subjected to agglomerative clustering in order to group similar acoustic segments. Finally, the orthography is iteratively aligned with the resulting transcription in terms of audio clusters in order to determine pronunciations of the training words. The approach is evaluated by applying it to two subsets of the TIMIT corpus, both of which have a closed vocabulary. It is found that, when vocabulary words occur often in the training set, the proposed technique delivers performance that is close to but lower than a system based on the TIMIT phonetic transcriptions. When vocabulary words are not repeated often in the training set, the best system is able to outperform its counterpart based on the TIMIT phonetic transcriptions, although recognition performance in both cases is poor.**

I. Introduction

We address the automatic generation of acoustic subword units and an associated pronunciation dictionary for speech recognition. Traditionally, the subword acoustic units and pronunciations used by a speech recogniser are phonetically-based and determined by a professional linguist. This procedure is extremely cumbersome and expensive. We propose to automatically generate both the acoustic units and the dictionary, using only the speech audio data and its orthographic transcription as input, as illustrated in Figure 1. If successful, this would increase the speed and reduce the cost of developing a speech recogniser for languages and accents for which resources in the form of pronunciation dictionaries and associated phone sets are not available, as is the case for many languages and accents in Southern Africa.

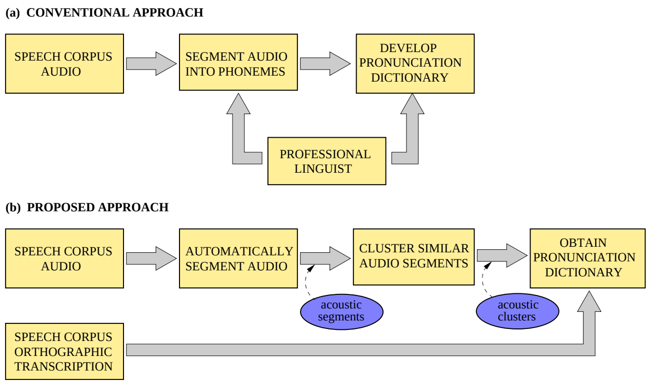
(a) CONVENTIONAL APPROACH
SPEECH CORPUS AUDIO
SEGMENT AUDIO INTO PHONEMES
DEVELOP PRONUNCIATION DICTIONARY
PROFESSIONAL LINGUIST
(b) PROPOSED APPROACH
SPEECH CORPUS AUDIO
AUTOMATICALLY SEGMENT AUDIO
CLUSTER SIMILAR AUDIO SEGMENTS
OBTAIN PRONUNCIATION DICTIONARY
acoustic segments
acoustic clusters
SPEECH CORPUS ORTHOGRAPHIC TRANSCRIPTION

Fig. 1. Comparison of (a) conventional and (b) proposed approaches to pronunciation dictionary generation.

Other research has also considered an iterative approach in which a set of acoustic models and automatically generated pronunciation dictionary are updated incrementally [1]. This system is bootstrapped from a configuration in which the graphemes of the words in the dictionary are used as subword units, and hence the alphabet used by the system defines the number of subword units. We will not make this assumption, but will attempt to find subword units by direct inspection of the audio training data. Nevertheless, the iterative approach proposed by [1] has influenced our work.

II. Data

For experimental evaluation we will make use of the TIMIT speech corpus, which contains a total of 6300 recorded sentences collected from eight different dialect regions in the United States. A total of 10 sentences were recorded for each of 630 different speakers, and there are male and female speakers from each dialect region. The recorded sentences can be divided into three categories:

- Phonetically diverse sentences (SI). These sentences were chosen from a large corpus of existing text to provide rich phonetic coverage and were designed to exploit the differences in the dialects.
- Phonetically compact sentences (SX). These sentences were designed by hand to provide a rich variety of phonetic segments and phonetic contexts.
- Unique sentences (SA). These two sentences were specially designed to demonstrate the effect of dialect on the acoustic characteristics of American English speech.

Sentence type	Sentences	Speakers	Utterances	Sents/spkr
Dialect (SA)	2	630	1260	2
Compact (SX)	450	7	3150	5
Diverse (SI)	1890	1	1890	3
Total	2342		6300	10

TABLE I
Composition of the TIMIT corpus.

Table I presents a breakdown of the composition of the TIMIT corpus. The table shows that only the SA sentences are repeated by every speaker. In addition every speaker reads five phonetically compact (SX) sentences, and each SX sentence is read by seven different speakers. Finally, each speaker also reads three phonetically diverse (SI) sentences, and each SI sentence is read by only one speaker.

Two subsets of the TIMIT corpus, shown in Table II, were used for the experimental evaluation of our approach.

- The SA sentences are used in isolation for testing and training. Due to the very small closed vocabulary and the high repetition rate of each word in the training data, this subset should provide a very optimistic scenario for our automatic subword unit determination methods.
- A subset of the SX and SI sentences was chosen such that all words in the respective test set are also present in the training set (i.e. there is a closed vocabulary). While all vocabulary words in the SA test set are repeated many times in the corresponding training set, the words in the SI+SX vocabulary are repeated with varying frequency in the corresponding training set. Hence the SI+SX subset represents a more realistic scenario in which to evaluate our automatic baseform determination methods.

TABLE II
TIMIT SUBSETS USED FOR EXPERIMENTATION.

	TIMIT subsets			
	SA		SI+SX	
	Train	Test	Train	Test
Vocabulary size (words)	21	21	2602	311
Number of utterances	924	48	1307	129
Number of speakers	462	24	450	93
Duration (minutes)	47.8	2.4	69.5	5.2

III. AUDIO SEGMENTATION

The first step in our procedure for the automatic generation of a subword representation is the segmentation of the audio into “phoneme-like” units. One approach to this problem is to consider the classification of the speech signal into voiced and unvoiced regions. This has been investigated by several authors for a variety of applications, including speech coding [2], [3] and speech recognition [4]. However, during preliminary tests, segmentation based on voicing seemed not to be a good basis for discovering subword units, since regions of sustained voicing led to very long segments.

A different family of algorithms tries to identify recurring phrases in unlabelled audio. These techniques are based on an alternative implementation of the dynamic time warping (DTW) algorithm, which allows it to detect local sub-matches between two audio segments [5]–[7]. These techniques are particularly suited to the detection of frequently recurring words or phrases in unlabelled audio from a single speaker and within a stable acoustic environment. However, they do not attempt to segment all the audio, but only to find frequently recurring sub-portions.

A few researchers have considered approaches that attempt to find segment boundaries in unlabelled audio by detecting points at which the time or spectral characteristics of the speech signal change strongly. Initial work located significant discontinuities in the speech spectra that had been subjected to a critical-band analysis [8]. Other authors have proposed variations on this technique [9], [10]. We have chosen the approach proposed in [11], which may be seen as a refinement on the work described in [8] leading to a more elegant algorithm requiring only a single user parameter [11], [12].

The first step in the segmentation procedure is to generate mel-frequency cepstral coefficient (MFCC) vectors from the audio. Twelve MFCCs, with the addition of log energy, first and second differential coefficients were used, resulting in a 39-dimensional feature vector. MFCC vectors are generated at a rate of one each 10ms and a window size of 20ms, corresponding to a half-frame overlap.

Next, a distance measure between consecutive feature vectors is defined in order to detect points at which the speech signal changes rapidly, and hence a segment boundary might be considered. The following measure is proposed by [11]:

$$d(v_1, v_2) = \arccos\left(\frac{v_1^t \cdot v_2}{\sqrt{(v_1^t \cdot v_1)\,(v_2^t \cdot v_2)}}\right)$$

where v_1 and v_2 are any two consecutive MFCC vectors, and $v_1^t \cdot v_2$ is the dot product between v_1 and v_2, such that:

$$v_1^t \cdot v_2 = \|v_1\|\,\|v_2\|\cos\theta$$

The distance $d(v_1, v_2)$ therefore corresponds to the angle between two consecutive vectors, and is used to segment the stream of MFCC vectors by means of the following criterion:

$$D(i) = \log\left(E\left(v_i\right)\right) \cdot d\left(\frac{v_{i-2} + v_{i-1}}{2}, \frac{v_{i+1} + v_{i+2}}{2}\right) > \delta \qquad (1)$$

This criterion states that the angle between two MFCC vectors will be weighted by the log energy of the current frame $E(v_i)$ in order to make a segmentation decision. Furthermore, the angle is calculated between the average of the two MFCC vectors preceding the current frame, v_{i-2} and v_{i-1}, and the average of the two following the current frame, v_{i+1} and v_{i+2}. The averages are used in order to take the variability of the angle over successive speech frames into account. Hence $D(i)$ will emphasise regions with strong changes in speech characteristics (large angle) and regions with high energy.

The audio is segmented by searching for all the peaks in $D(i)$ above a certain threshold. However in practice it is found that $D(i)$ has many minor peaks. Hence $D(i)$ is smoothed using a nine-point Hanning window, arranged symmetrically around v_i, as also proposed by [11]. The result of this audio segmentation process is a sequence of points in time at which the segment boundaries have been hypothesised.

IV. CLUSTERING

In order to determine how similar two acoustic segments of unequal lengths are, dynamic time warping (DTW) was used. This algorithm can be considered an application of dynamic programming, where the goal is to find the optimal alignment between two sequences, given some constraints. The result is the best frame-by-frame alignment between two acoustic sequences, as well as an overall score which can be used to quantify the quality of the alignment.

In order to group similar acoustic segments, we will make use of agglomerative hierarchical clustering. This clustering approach requires only the similarities between the units to be clustered to be known [13]. Initially, all acoustic segments are considered individual clusters. These are subsequently merged successively in an iterative fashion. The average similarity between all possible pairs of acoustic segments is used as a measure of cluster similarity.

V. PRONUNCIATION DICTIONARY GENERATION

After audio segmentation and clustering, we are left with a sequence of automatically generated acoustic clusters and associated orthographic transcription (word sequence). To generate a pronunciation dictionary, we proceed by first finding an initial alignment between these two sequences. From this alignment, a mapping is constructed between each word in the orthographic transcription and its respective sequence of acoustic clusters. This mapping is considered the initial dictionary, and is subsequently refined by iterative re-alignment.

A. Initial pronunciation dictionary

An initial alignment, which is a crude guess of how the acoustic clusters align with the words in the orthographic transcription, is needed to begin the dictionary generation process. We obtain this initial alignment by counting the number of acoustic clusters and the number of words in each training utterance. The acoustic clusters are aligned with the words while trying to keep the number of acoustic clusters per word approximately constant. There are alternatives to this naïve initialisation approach, such as trying to take the lengths of the words into account. However time did not permit these to be explored.

B. Intermediate pronunciation dictionary

Since both the sequence of words and acoustic clusters occur in a fixed order, an HMM can be used as an appropriate statistical model with which to perform their alignment. The states of each HMM correspond to the words and the sequence of acoustic clusters to the observations of the HMM. Each utterance can then be represented by an HMM consisting of a sequence of word HMMs. Finding the optimal alignment is reduced to finding the state sequence that maximises the probability of the observation sequence, which can be achieved by means of the Viterbi algorithm. In order to model the sequential nature of the orthography, a left-to-right HMM structure with no skips was chosen. Each node of each HMM has associated with it a probability distribution, describing how likely it is for the state to be associated with each acoustic cluster. We will estimate these observation probabilities from a relative frequency that is obtained from the most recent alignment. In particular, the probability that cluster c_i is associated with word w_j is calculated as:

$$P(c_i|w_j) = \frac{\text{Number of times } c_i \text{ is aligned with } w_j}{\text{Total number of clusters aligned with } w_j}$$

The totals on the right hand side of the above equation are obtained by accumulating the counts obtained from the most recent alignment of the entire training set with the corresponding acoustic cluster sequences. The dictionary obtained from this improved alignment is referred to as the intermediate dictionary.

C. Final pronunciation dictionary

The intermediate dictionary will generally have many different pronunciations for the same word. Some of these pronunciations may rarely be associated with a word. So, in a last step, we aim to prune out such infrequent candidates from the dictionary.

First the intermediate dictionary is used to create initial acoustic models using the HTK tools, as illustrated by steps one and two in Figure 2. Single-mixture flat-start context-independent HMMs are initialised from the global mean and variance of the training set audio for each acoustic cluster. These initial models are then updated by performing five iterations of embedded re-estimation. In each case each HMM consists of three states arranged in a left-to-right topology, with a single Gaussian mixture per state. Acoustic observations are parameterised as MFCCs, with appended energy, first and second differentials.

The HMM models obtained in the above process, together with the intermediate dictionary, are then used to perform a forced alignment between the orthography and the acoustic data, using all the pronunciations variants contained in the dictionary (step three in Figure 2). In this process, pronunciation variants are presented in parallel to the Viterbi decoder. Not all pronunciation variants will be favoured by the forced alignment, and those rarely used are subsequently pruned from the dictionary in step five of Figure 2. The most recently updated HMMs as well as the new dictionary are then used to perform further forced alignments, followed by updates to the HMMs as shown in steps six, seven and eight of Figure 2. This process repeats until the dictionary no longer changes between successive iterations. Finally, the number of Gaussian mixtures per HMM state is increased from one to two and then four, six and eight where each increase is followed by a further set of five Baum-Welch re-estimation iterations.

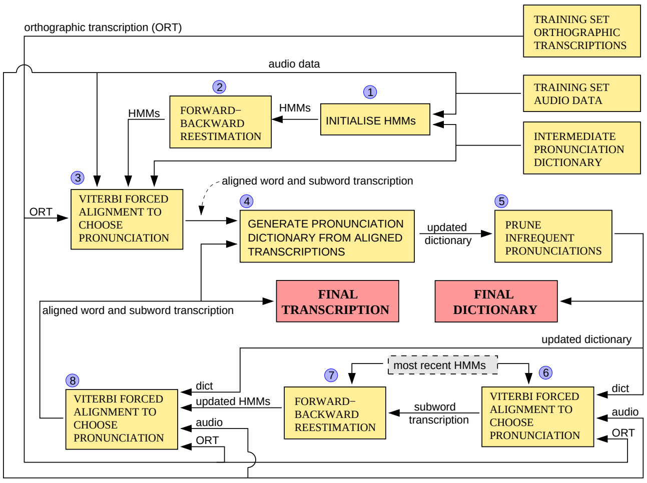
TRAINING SET ORTHOGRAPHIC TRANSCRIPTIONS
orthographic transcription (ORT)
audio data
TRAINING SET AUDIO DATA
INTERMEDIATE PRONUNCIATION DICTIONARY
1
INITIALISE HMMs
HMMs
2
FORWARD–BACKWARD REESTIMATION
HMMs
3
VITERBI FORCED ALIGNMENT TO CHOOSE PRONUNCIATION
ORT
aligned word and subword transcription
4
GENERATE PRONUNCIATION DICTIONARY FROM ALIGNED TRANSCRIPTIONS
updated dictionary
5
PRUNE INFREQUENT PRONUNCIATIONS
aligned word and subword transcription
FINAL TRANSCRIPTION
FINAL DICTIONARY
updated dictionary
most recent HMMs
8
VITERBI FORCED ALIGNMENT TO CHOOSE PRONUNCIATION
dict
updated HMMs
audio
ORT
7
FORWARD–BACKWARD REESTIMATION
subword transcription
6
VITERBI FORCED ALIGNMENT TO CHOOSE PRONUNCIATION
dict
audio
ORT

Fig. 2. Generation of the final dictionary.

The pruning algorithm (step five in Figure 2) is governed by a pruning threshold whose optimal value must be determined experimentally. First, all the unique sequences of acoustic clusters associated with each particular word are counted (step four in Figure 2). The result indicates the probability of each pronunciation in the current alignment. Starting with the highest probability pronunciation, probabilities are accumulated until the total exceeds the pruning threshold. All pronunciations that form part of this accumulated total are retained, while the remainder are pruned from the dictionary. However, at least one pronunciation is retained for each word. The result is that the word-to-cluster sequence mappings that occur often in the training set alignment are retained, while infrequent ones are not. The final dictionary produced by this process can be used to train a new set of acoustic models that can be used in a speech recognition system.

VI. Experimental Results

The first set of experiments was performed using the SA sentences for both training and testing, as described in Section II. The purpose of these experiments was to determine how the proposed approach would perform when presented with input data that is highly repetitive. The highly repetitive orthography of the SA sentences will provide ample data for each word and hence represents a very optimistic scenario.

A. Segmentation threshold

The average length and therefore also the overall number of acoustic segments can be varied by varying the segmentation threshold, as described in Section III.

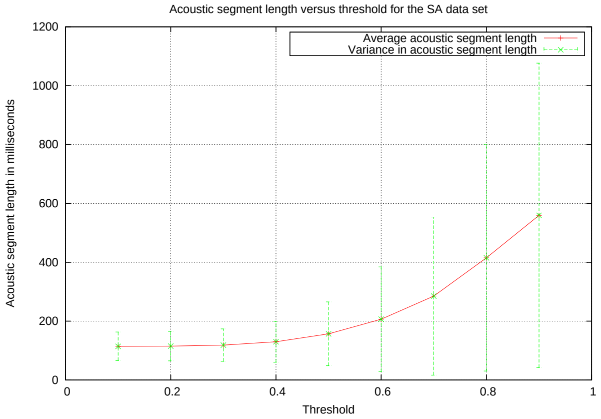

Fig. 3. Acoustic segment length as a function of the segmentation threshold for the SA data set. The standard deviation is indicated by error bars.

Figure 3 shows how the average segment length changes as a function of the segmentation threshold for the SA training set. As a point of reference, we consider the average length of the phonemes in the TIMIT phonetic transcriptions is 110ms We conclude that a low threshold, in the region of 0.2, would be a good value for experimentation.

B. Experiments using the SA data

Speech recognition results for the SA data set, using a segmentation threshold of 0.2, are presented in Figure 4 for eight mixture HMMs. This graph indicates the performance of systems trained using automatically-determined clusters, as well as phonetic clusters (i.e. the TIMIT phonetic transcriptions). The latter serves as a baseline to the former. Systems using 32, 64, 128, 256 and 512 clusters were considered.

The dictionary pruning threshold was varied between a value of 0.1 and 0.9, to produce a variety of systems, whose performance in terms of speech recognition accuracy (word accuracy) are shown in Figure 4. Speech recognition was performed using HTK, which performs Viterbi decoding on the audio input using the set of HMMs and the dictionary produced by the process illustrated in Figure 2. A word-loop grammar, in which all 21 vocabulary words are equally likely, is used for speech recognition. We see that the phonetic segmentation led to the highest recognition accuracy in almost all cases. This indicates that the pre-defined phonemes in the TIMIT corpus represent the best clusters, and that the automatic segmentation does not perform as well as the phonetic segmentation.

The results in Figure 4 also indicate the effect of varying the number of clusters produced by the clustering stage. Graphs are shown for systems based on 32, 64, 128, 256 and 512 clusters. These numbers refer to the number of clusters produced by the clustering stage. However, during dictionary refinement, the pronunciation pruning algorithm discards pronunciations, which generally also leads to a reduction in the number of subword units used by the final system. Table III indicates how many different subword units remained in the final dictionary for each user specified number of clusters.

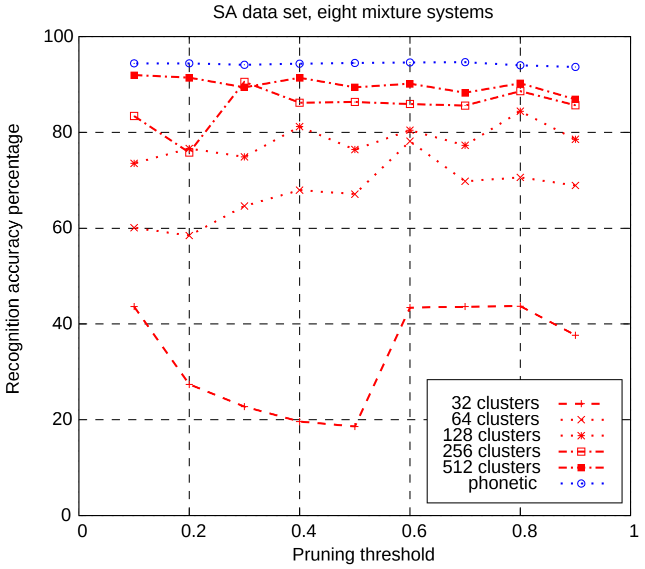

Fig. 4. Performance of 8-mixture systems developed using the SA data set.

In initial dictionary	In final dictionary
512	32
256	26
128	19
64	14
32	13

TABLE III
INITIAL VERSUS FINAL NUMBER OF SUBWORD UNITS FOR THE SA EXPERIMENTS.

The table shows that, when the clustering stage produces 512 clusters, only 32 of these remain in the final dictionary. This is comparable to the 48 phonemes defined by TIMIT and leads to the best performing system. The small number of remaining clusters is probably related to the limited vocabulary of the SA data set.

From the results in Figure 4, it is clear that systems with fewer than 256 clusters perform significantly worse than the baseline system using the time phonetic annotations. Secondly, although the baseline was the best performing system, the automatically-determined system using 512 clusters achieved performance which was almost as good.

C. Experiments using the SI+SX data

The second set of experiments was performed using the SI+SX data set, as described in Section II. This data is more challenging because the sentences are not repeated among the speakers, as they were for the SA data. Furthermore, the SI+SX data set has a larger vocabulary (2595 words) and not all the test words are repeated multiple times in the training data. The scenario makes it considerably more difficult to obtain reliable pronunciations and to train good HMMs for recognition. In particular, even the system trained on the TIMIT phonetic transcriptions has a low recognition accuracy.

The same segmentation threshold of 0.2 used for the SA experiments was employed again for the SI+SX data set. The results for an eight mixture system are presented in Figure 5. As for the SA experiments, a word-loop grammar, in which any of the 2595 vocabulary words may follow each other with equal probability, was used during decoding.

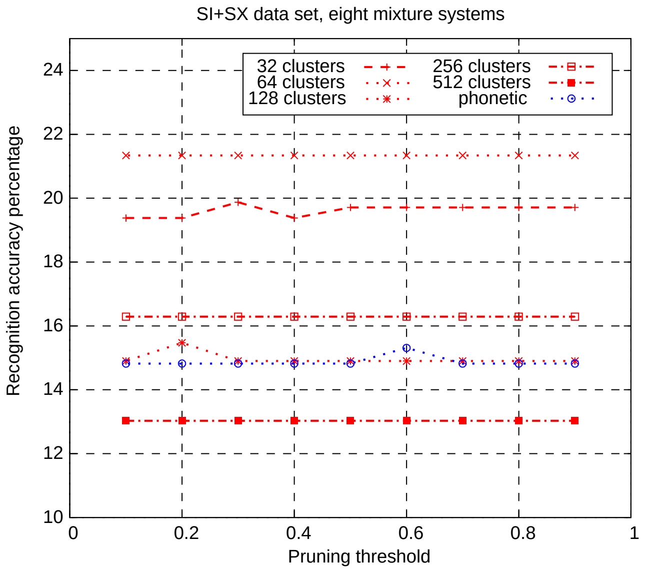

Fig. 5. Results of SI+SX data set using eight mixture systems.

From the results we see that the performance of all systems is poor. In general, a recognition accuracy of around 20 percent was achieved. The recognition accuracy curves are also fairly flat compared to those obtained for the SA data set. Since multiple repetitions and hence pronunciations of each vocabulary word do not often occur, the dictionary pruning algorithm eliminates most pronunciation sequences even at a threshold of 0.2.

In initial dictionary	In final dictionary
512	341
256	177
128	93
64	46
32	27

TABLE IV
INITIAL VERSUS FINAL NUMBER OF SUBWORD UNITS FOR THE SI+SX EXPERIMENTS.

Table IV indicates how many different subword units remain in the final dictionary for each user-specified number of clusters. The table shows that, when the clustering stage produces 512 clusters, 341 remain after dictionary generation. For 32 user-specified clusters, the dictionary generation stage retains almost all of them. For the case of 64 user-specified clusters, 46 remained after the dictionary generation stage. This number of clusters is close the number of defined phonemes in TIMIT and the associated system also achieves the best results overall. For 128 clusters and more, an increasing number are eliminated during the iterative refinement of the final dictionary.

D. Interpretation of results

For the SA experiments, the performance of the systems improved as the number of clusters increased. Furthermore, it was seen that the overwhelming majority of clusters were pruned when generating the final dictionary. The same was not true for the SI+SX system, for which best performance was achieved for an intermediate number of clusters (64), and a much smaller proportion of clusters were pruned during dictionary generation.

A major difference between the SA and the SI+SX data sets is the frequency of repetition of the training words in the training sets. The high level of repetition in the SA set allowed many postulated pronunciations to compete during dictionary generation. This led to a greater degree of dictionary pruning, and to a better performing system. However the acoustic models based on the phonetic transcription still led to the best results for the SA data set.

For the SI+SX data set, however, most training words occur only once. This means there is far less opportunity for competing pronunciations to be generated, and for the subsequent dictionary pruning to remove less well-performing candidates. Nevertheless, acoustic models based on the best configuration (64 clusters) led to better recognition accuracies than models based on the phonetic transcriptions. It should be borne in mind, however, that the recognition accuracies were very low overall.

VII. Summary and Discussion

The success of the clustering of acoustic segments to discover subword units depends critically on the quality of the segments themselves. Furthermore, the quality of the segments and resultant clusters critically affect the quality and ultimate success of the dictionary. By listening to a random selection of the audio segments produced by our system, it was concluded that although sometimes the segments appeared to be plausible subword units, sometimes they were not. Badly-formed segments most certainly have had a detrimental effect on the success of the overall approach. However time did not permit this issue to be investigated with more rigour.

While the segmentation and clustering stages were based on known and published approaches, our procedure for the automatic determination of a pronunciation dictionary is, as far as we know, new. The dictionary is determined by an iterative process of alignment between the automatically determined subword transcription, the associated orthographic transcription, and the corresponding audio data. Subword and orthographic transcriptions are aligned by modelling the orthography as a hidden Markov model (HMM), where the subword units are the observations and the states of the HMM are the words. The resulting dictionary is used to align the possible pronunciations with the audio, and thereby discard poorly matching pronunciation variants. The process is iterated until some degree of convergence is achieved. When this dictionary generation process is presented with the true TIMIT phonetic transcriptions, instead of the automatically determined subword units, a pronunciation dictionary containing reasonable pronunciations (according to informal inspection) was determined. When presented with the automatically determined subword units, the dictionary was difficult to analyse, but nevertheless lead to a working system.

The overall approach was evaluated by testing it using two subsets of the TIMIT corpus. The first, termed the “SA” system, used the two phonetically rich sentences repeated by every speaker as training and testing material. Although there is no speaker overlap between the test and train sets, the same two word sequences constituted both. This is a very optimistic testing scenario, since the vocabulary is small (21 words) and each word in the training set is repeated many times (462 times), albeit by different speakers. In this testing scenario, it was found that a system trained on automatically-determined subword units could achieve a performance nearly as good as one trained on the true TIMIT phonetic transcriptions. Due to the small vocabulary and the high repetition rate of training words, the word accuracies achieved by these systems were rather high.

The second subset of the TIMIT corpus was drawn from the SI and the SX sentences, which are also phonetically rich, but which are repeated by only seven speakers or only once for the SX and SI sentences respectively. The training and testing subsets were chosen in such a way as to ensure a closed vocabulary, i.e. that each word in the test set also occurs at least once in the training set. Systems trained on this SI+SX subset exhibited much poorer performance than those based on the SA subsets. However, the best systems based on an automatically determined subword units were able to outperform those based on the TIMIT phonetic transcriptions. Notwithstanding the low word accuracies, this is a very promising result.

VIII. Conclusion

It was possible to obtain working speech recognition systems using only the orthographic transcriptions and the audio data of the training set as input. In particular, no pronunciation dictionary or other subword information was employed. The overall system is complex, and time did no permit thorough testing and analysis. However, the limited test results are rather positive. If the techniques described and proposed in this paper can be further analysed and refined, it could be an important step for the development of speech recognition systems for under-resourced languages or dialects, for which the extensive phonetic resources conventionally required, are not available.

Acknowledgements

Parts of this work were executed using the High Performance Computer (HPC) facility at Stellenbosch University.

References

[1] R. Singh, B. Raj, and R. M. Stern, "Automatic generation of subword units for speech recognition systems," *IEEE Transactions on Speech and Audio Processing*, vol. 10, no. 2, pp. 89–99, 2002.

[2] H. Tolba and D. O'Shaugnessy, "Robust automatic continuous speech recognition based on a voiced-unvoiced decision," in *Proceedings of ICSLP*, 1998.

[3] S. Ahmadi and A. Spanias, "Cepstrum-based pitch detection using a new statistical v/uv classification algorithm," *IEEE Transactions on Speech and Audio Processing*, vol. 7, no. 3, pp. 333–338, 1999.

[4] A. Zolnay, R. Schlter, and H. Ney, "Extraction methods of voicing feature for robust speech recognition," in *Proceedings of Eurospeech*, 2003.

[5] A. Park and J. Glass, "Towards unsupervised pattern discovery in speech," in *Proceedings of ASRU*, 2005.

[6] M. Gajjar, R. Govindarajan, and T. Sreenivas, "Online unsupervised pattern discovery in speech using parallelization," in *Proceedings of Interspeech*, 2008.

[7] A. Muscariello, G. Gravier, and F. Bimbot, "Audio keyword extraction by unsupervised word discovery," in *Proceedings of Interspeech*, 2009.

[8] G. Aversano, A. Esposito, and M. Marinaro, "A new text-independent method for phoneme segmentation," in *Proceedings of 44th IEEE Midwest Symposium on Circuits and Systems*, 2001.

[9] H. Murthy and V. Gadde, "The modified group delay function and its application to phoneme recognition," in *Proceedings of ICASSP*, 2003.

[10] L. Golipour and D. O'Shaughnessy, "A new approach for phoneme segmentation of speech signals," *Proceedings of InterSpeech*, 2007.

[11] L. ten Bosch and B. Cranen, "A computational model for unsupervised word discovery," in *Proceedings of Interspeech*, 2007.

[12] L. ten Bosch, H. Baayen, and M. Ernestus, "On speech variation and word type differentiation by articulatory feature representations," in *Proceedings of ICSLP*, 2006.

[13] S. B. Everitt, *Cluster Analysis*, 3rd ed. Cambridge University Press, 1993.