LONGER-LENGTH ACOUSTIC UNITS FOR CONTINUOUS SPEECH RECOGNITION

Annika Hämäläinen, Johan de Veth, and Lou Boves

Department of Language and Speech, Radboud University Nijmegen
P.O. Box 9103, 6500 HD Nijmegen, The Netherlands
phone: +31 (0)24 361 57 64, fax: +31 (0)24 361 29 07, email: {A.Hamalainen, J.deVeth, L.Boves}@let.ru.nl
web: http://lands.let.ru.nl

ABSTRACT

Recent research on the TIMIT database suggests that longer-length acoustic units are better suited for modelling pronunciation variation and long-term temporal dependencies in speech than traditional phoneme-length units, yielding substantial improvements in recognition accuracy [9]. In this paper, we investigate whether similar improvements can be gained on another database, viz. excerpts from novels in a Dutch library for the blind. We use a hierarchical method that employs a mixture of word-, syllable- and phoneme-length units. Our results show that the approach does increase the word accuracy, but to a lesser extent than expected. The paper discusses possible explanations for the finding.

1. INTRODUCTION

Large-vocabulary continuous speech recognition (LVCSR) systems conventionally use context-dependent phone models, such as triphones, to model the elementary acoustic units of speech. The main advantage of triphones is that the fixed number of phonemes in a given language guarantees the robust training of acoustic models when reasonable amounts of training data are available and when state tying methods are used to deal with triphones with insufficient training data. To some extent, triphones are able to model short-term contextual effects that cause variations in the way a particular phoneme is produced; according to [1], phoneme substitution and reduction are well captured by triphones.

However, the use of triphones as the elementary units for speech sound modelling can be called into question because of the complexities of speech that triphones cannot capture. First, coarticulation effects typically have a long time span, and the corresponding spectral and temporal dependencies are not easy to capture in triphones, which model speech segments with very limited a duration [2]. Second, the use of triphones is based on a simplified view of speech where words are represented as sequences of discrete phonemes ('beads on a string') [3]. Within this framework, pronunciation variation can only be represented in terms of phoneme-level substitutions, deletions and insertions. Moreover, this description has a limited capability of making effective use of any higher-level dependencies related to e.g. syllable structure. These inherent restrictions of triphones raise the question how the long-term spectral and temporal dependencies present in natural speech could be modelled in LVCSR.

We believe that the solution lies in the use of longer-length acoustic units that have the spectral and temporal dependencies embedded into them. Part of the motivation for this comes from studies of human speech production and recognition. For example, [4] claims that humans rely on an episodic memory, which stores details of spectro-temporal patterns observed earlier. Although it remains an open question which longer-length acoustic units are represented in an episodic memory, these units are most likely (much) longer than a phoneme.

The most obvious candidates for longer-length acoustic units are the word and the syllable. In particular, the use of the syllable as an elementary unit of speech is supported by studies of human speech production and perception [e.g. 5, 6]. For the purpose of automatic speech recognition (ASR), particularly attractive features of the syllable are regularities within syllables, as well as their low deletion rate. Based on analyses of the hand-labelled Switchboard corpus of American English spontaneous speech, [7] reports that syllable onsets generally maintain their canonical form, while the coda elements often get deleted or assimilated with the onset of the following syllable, and the nuclei remain but often change quality. [7] also reports the deletion rate of syllables to be just 1%, compared with the 12% deletion rate of phonemes. For these reasons, it seems that a representation of words as a sequence of syllables can cover a substantial amount of pronunciation variation with fewer alternatives in the lexicon than when a phonemic representation is used.

The use of longer-length acoustic units in ASR is not a novel idea: word models are commonly used in applications with limited vocabularies, such as digit recognition, and command and control applications. The use of syllable [e.g. 2, 8-10, 12], demi-syllable [11] and multiphone [12] models has also been suggested before, but the problem with the use of longer-length acoustic units in LVCSR has been the sparsity of training data when training models from scratch. When trained without clever data sharing, longer-length acoustic units require more training data than phoneme-based units [9]. As the units become longer, the number of units with little or no acoustic data available for model parameter estimation will increase. If the longer-length units are words, there is an unbounded increase in the number of possible units.

Solutions to the training data sparsity problem have been suggested in [9], and more recently in [12]. Due to a surprisingly low baseline, the performance gain reported in [9] may not be fully representative. Nevertheless, since the approach has high face validity, it was taken as the starting point for our own research. The approach is summarised in Subsection 3.3.

Our study investigates the degree to which the method suggested in [9] is capable of modelling speech of another language and reading style than carefully read American English. For this reason, we implemented the method suggested in [9],

verified that our implementation was valid for American English, and then tested the method for Dutch speech read in a more lively style.

This paper is further organised as follows. The speech material is described in Section 2. The experimental set-up (feature extraction, language modelling, and acoustic unit definition) is detailed in Section 3. The results are presented and discussed in Section 4. Finally, our conclusions are formulated in Section 5.

2. SPEECH MATERIAL

The American English speech material originated from the TIMIT database [13], and the Dutch speech material was extracted from the Spoken Dutch Corpus (Corpus Gesproken Nederlands; CGN) [14]. TIMIT contains carefully read speech, which is manually labelled and includes time-aligned, manually verified phonetic and word segmentations. For this study, the original set of labels was reduced to a set of 45 phone labels. The CGN material used in this study was read speech from the library for the blind. It comprises manually verified phonetic and word labels, as well as manually verified word-level segmentations. A set of 37 phone labels was used for CGN. Apart from differences in the labelling and segmentation protocol, there are two important differences between the two sets of read speech. Firstly, TIMIT consists of sentences read in isolation, whereas the CGN data consist of coherent fragments of text at least the size of a paragraph. Secondly and more significantly, due to the readers' involvement in the stories, the CGN texts are read in a more lively style than the TIMIT sentences.

The data for each language were divided into three sets: a training set for the training of the acoustic models, a test set for the evaluation of the acoustic models, and a development test set for the optimisation of the following training/recognition parameters: the minimum number of tokens required to robustly train longer-length acoustic units, the language model scaling factor, and the word insertion penalty. None of the data had a high background noise level. Details of the data sets are presented in Tables 1 and 2.

Table 1: TIMIT data sets.

	Training	Test	Devel.	Total
Ortographic word tokens	30,132	9,455	1,570	41,157
Speakers/ Female/Male	462/ 136/326	144/ 48/96	24/ 8/16	630/ 192/438
Duration (hh:mm:ss)	03:08:42	00:59:13	00:09:43	04:17:38

Table 2: CGN data sets.

	Training	Test	Devel.	Total
Ortographic word tokens	45,172	7,917	7,507	60,596
Speakers/ Female/Male	125/ 70/55	125/ 70/55	125/ 70/55	125/ 70/55
Duration (hh:mm:ss)	04:51:27	00:51:34	00:48:13	06:31:14

3. EXPERIMENTAL SET-UP

3.1 Feature Extraction

The feature extraction was carried out at a frame rate of 10 ms, applying a pre-emphasis of 0.97. 12 Mel Frequency Cepstral Coefficients (MFCCs) and log-energy with corresponding first and second order time derivatives were calculated for each recording, resulting in a total of 39 features. Channel normalisation was applied by means of cepstral mean normalisation over individual sentences for TIMIT and complete recordings (with a mean duration of 3.5 minutes) for CGN. For training and testing purposes, the CGN data were chunked to sentence-length entities. The feature extraction was performed using HTK [15].

3.2 Lexica and Language Models

The recognition lexicon and word-level bigram network for each language were built using all orthographic words in the training and test sets. In effect, with the language models and lexica used, no out-of-vocabulary words appeared in the tasks. The vocabulary consisted of about 6,000 words in the case of the TIMIT data and 10,500 words in the case of the CGN data. The test set perplexity was 16 for TIMIT and 46 for CGN. The test set perplexities were computed on a per-sentence basis using HTK [15].

3.3 Acoustic Modelling

In preparation for building mixed-unit recognisers that employed a selection of word-, syllable- and phoneme-length units, three different types of recognisers were built for each language: a triphone recogniser, a word-unit recogniser and a syllable-unit recogniser. As a sanity check, the implemented procedures were tested on the TIMIT data before using them for the CGN data. The baseline performance for each language was determined by the performance of the triphone recogniser.

A standard procedure with decision tree –based state tying was used to train the triphone recognisers [e.g. 15]. As opposed to [9], which used a flat start Baum-Welch re-estimation strategy, the TIMIT triphones were bootstrapped using the manually verified phonetic segmentations of the sentences. For CGN, 32-Gaussian monophones were first bootstrapped using linear segmentation within the manually verified word segments. The monophones were used to perform a forced alignment of the training data; the CGN triphones were bootstrapped using the resulting phone segmentations.

The principles of building the word-, syllable- and mixed-unit recognisers were implemented as described in [9]. To summarise, the context-free word and syllable units of the word- and syllable-unit recognisers were initialised using the model parameters of the 8-Gaussian triphone models corresponding to the underlying (canonical) phonemes of the words and syllables, respectively. To embed the spectral and temporal dependencies into them, the word and syllable units which appeared frequently enough in the training data were trained further using the Baum-Welch re-estimation algorithm. The optimal frequencies were determined empirically, by varying the value for the minimum number of tokens required for the further training of a unit, and monitoring the recognition performance achieved on the development test set. Only robustly trained word and syllable units from the word- and syllable-unit recognisers were used in the mixed-unit recogniser; triphones were backed off to in the case of words and syllables that did not occur frequently enough in

the training data to allow the robust training of corresponding word and syllable units.

Two types of mixed-unit recognisers were built: one that corresponded to the mixed-unit recogniser of [9] in that it comprised multisyllabic word units, syllable units – including monosyllabic words – and triphones (mixed-unit recogniser A), and another that comprised syllable units – including monosyllabic words – and triphones (mixed-unit recogniser B).

4. RESULTS AND DISCUSSION

4.1 TIMIT

The results we obtained for the TIMIT data are presented in Table 3 (3rd column), together with the results reported in [9] (2nd column). Our triphone results are for triphones with 16 Gaussian mixture components (best performing triphones); the results of [9] are for triphones with 8 Gaussian mixture components. Mixed-unit recogniser A contained 1 word unit and 151 syllable units, whereas mixed-unit recogniser B contained 151 syllable units.

As can be seen in Table 3, the use of longer-length acoustic units resulted in substantial gains in word accuracy in both studies. While the absolute values differ between the two studies, the ranking of the different recognisers with respect to word accuracies are the same: mixed-unit recogniser A performs the best, followed by the word- and syllable-unit recognisers. The word accuracy of mixed-unit recogniser B, which does not contain multisyllabic words, is exactly the same as that of mixed-unit recogniser A. Performances of a similar level could be expected, as the only difference between these two recognisers was the modelling of the bi-syllabic function word "every"; in mixed-unit recogniser B it was modelled using two context-free syllable units, whereas in mixed-unit recogniser A it was modelled using a context-free word unit.

We have been unable to determine why the baseline performance from which we started was so much higher than the baseline performance reported in [9]. There are at least two possibilities. First, we used the manually verified phonetic segmentations to initialise the acoustic model training (instead of a flat start); this may have resulted in better triphone models. Second, our language model may have been much more powerful. Yet, our results are similar to [9] in a *qualitative* way. We take this as proof of our implementation of the longer-length unit recognisers being valid.

Although we were not able to reproduce the enormous performance gain reported in [9], we still obtained an impressive 42% relative reduction in word error rate. Thus, it appears that the use of a mixed-unit recognition scheme does indeed yield a very large performance gain for TIMIT.

Table 3 Word accuracies achieved on TIMIT in [9] and in our study, and word accuracies achieved on CGN. Our results are presented with a 95% confidence interval.

Recogniser type	TIMIT [9]	TIMIT	CGN
Triphone	74	91.9 ± 0.6	91.8 ± 0.6
Word-unit	87	94.0 ± 0.5	92.9 ± 0.6
Syllable-unit	85	93.5 ± 0.5	92.9 ± 0.6
Mixed-unit A	90	95.3 ± 0.5	N/A
Mixed-unit B	N/A	95.3 ± 0.5	93.3 ± 0.6

4.2 CGN

The recognition results on the CGN data are presented in Table 3 (4th column). The triphone results were achieved using triphones with 8 Gaussian mixture components (best performing triphones). There were no multisyllabic word units in mixed-unit recogniser A; mixed-unit recogniser A was identical to mixed-unit recogniser B. Mixed-unit recogniser B contained 94 syllable units.

In the case of the triphone recogniser, the results we obtained for the CGN data were of the same level as the results we reached for the TIMIT data – regardless of the large difference in the test set perplexities. This suggests that the acoustic perplexity of the CGN data is much lower than in TIMIT. However, similar improvements were not reached when moving on to the word-, syllable- and mixed-unit recognisers. They all perform significantly better than the triphone recogniser, but there are no significant differences between their performances – unlike in the case of the TIMIT data.

Although the performance gain obtained for the CGN data is obviously much smaller than the 42% reduction in word error rate for TIMIT, we did obtain a substantial improvement: the relative reduction in word error rate was 18%. Other studies applying a mixed-unit recognition scheme have also failed to reach the kind of improvements gained on TIMIT. The absolute improvement in recognition accuracy obtained in [10] was only 0.5%. In [12] the performance gain due to the inclusion of longer-length acoustic units depended heavily on the recognition task: for telephone numbers, the performance even decreased. Thus, it seems that TIMIT is special in respects that enhance the impact of improved acoustic modelling. For the time being, we can only advance several hypotheses – most of which remain to be extensively tested – about the reasons for performance differences between TIMIT and CGN.

4.3 What Makes TIMIT Different from CGN?

This section discusses possible explanations for the large performance differences between TIMIT and other corpora when a mixed-unit recognition scheme is employed. Since we encountered such differences when carrying out experiments on the CGN data, the discussion is focused on possible differences between the two corpora.

One hypothesis – suggested by the different number of words per minute between TIMIT and CGN – can be rejected, viz. that differences in the word structure of the two data sets might account for the differences in performance. Long, polysyllabic words are easier to recognise than short words; if a large proportion of words in the data were long, polysyllabic words, improved acoustic modelling could not add much to the recognition performance. To establish whether the CGN data had a higher proportion of polysyllabic words than the TIMIT data, the proportion of word tokens containing different numbers of syllables was calculated. The word structure of the TIMIT and CGN data is illustrated in Table 4. The figures are remarkably similar considering that we are looking at two distinct, albeit Germanic, languages. However, in interpreting the figures in Table 4, it should be taken into account that the further trained syllables in TIMIT cover less of the test data (48.9%) than the further trained syllables in CGN (56.8%) – despite the fact that there were 1.6 times more further trained syllable units for TIMIT. This suggests that the amount of pronunciation variation in CGN is much larger.

Table 4 Word structure of TIMIT and CGN data.

Number of syllables	TIMIT	CGN
1	63.1%	62.2%
2	22.7%	22.6%
3	9.3%	9.9%
4	3.5%	3.9%
5	1.2%	1.1%
≥ 6	0.2%	0.3%

The hypothesis that a relatively higher degree of pronunciation variation might be present in the CGN data as compared with TIMIT – even though the number of speakers in the CGN data is smaller than in TIMIT – can be justified. The CGN data originate from the library for the blind, and it becomes evident from listening to the speech that, due to the readers' involvement in the stories, the reading style is much more lively than that of TIMIT. The higher the degree of pronunciation variation in the data is, the more tokens are needed to train robust longer-length acoustic units that are able to generalise to pronunciation variation in unseen data. The hypothesis of a higher degree of pronunciation variation in the CGN data is supported by our experiments with the minimum number of tokens required for the robust training of longer-length acoustic units: in the case of the CGN syllable units, a minimum of 130 training tokens was deemed optimal, compared with 60 for the TIMIT syllable units. Counting the number of different pronunciations of each word/syllable in the corpus would be another method to analyse the role of pronunciation variation. However, CGN comes with broad phonetic transcriptions, while the TIMIT transcriptions are more detailed and probably also more accurate. Therefore, the current data do not enable a direct comparison of the numbers of pronunciation variants.

Additional experiments are under way to determine what makes TIMIT special in terms of the effects of improved acoustic modelling. We expect that the results of these experiments will enable us to more specifically establish in which conditions the use of longer-length acoustic units can be expected to yield a significant advantage over the use of triphones.

5. CONCLUSIONS

In this paper, recognition results obtained using longer-length acoustic units for Dutch read speech were presented. These results were contrasted with the recognition results for similar experiments on American English read speech. In the case of both languages, significant improvements over the performance of a triphone recogniser were obtained when using a mixed-unit recogniser comprising word-, syllable- and phoneme-length acoustic units.

6. ACKNOWLEDGEMENTS

The research was carried out within the framework of the Interactive Multimodal Information eXtraction (IMIX) program, which is sponsored by Netherlands Organisation for Scientific Research (NWO).

REFERENCES

[1] D. Jurafsky, W. Ward, Z. Jianping, K. Herold, Y. Xiuyang, and Z. Sen, "What kind of pronunciation variation is hard for triphones to model?" in *Proc. ICASSP-2001*, Salt Lake City, Utah, USA, May 8-11. 2001, vol. I, pp. 577-580.

[2] A. Ganapathiraju, J. Hamaker, M. Ordowski, G. Doddington, and J. Picone, "Syllable-based large vocabulary continuous speech recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 9(4), pp. 358-366, May 2001.

[3] M. Ostendorf, "Moving beyond the 'beads-on-a-string' model of speech," in *Proc. IEEE ASRU-99*, Keystone, Colorado, USA. Dec 12-15, 1999.

[4] S. Goldinger, "Echoes of echoes? An episodic theory of lexical access," *Psychological Review*, vol. 105(2), pp. 251-279, Apr 1998.

[5] N.O. Schiller, A.S. Meyer, and W.J.M. Levelt, "The syllabic structure of spoken words: Evidence from the syllabification of intervocalic consonants," *Language & Speech*, vol. 40, pp. 103-140, 1997.

[6] C. Pallier, "Phonemes and syllables in speech perception: size of attentional focus in French," in *Proc. Eurospeech-97*, Rhodes, Greece, Sept 22-25. 1997, vol. 4, pp. 2159-2162.

[7] S. Greenberg, "Speaking in shorthand – A syllable-centric perspective for understanding pronunciation variation," *Speech Communication*, vol. 29, pp. 159-176, 1999.

[8] R.J. Jones, S. Downey, and J.S. Mason, "Continuous speech recognition using syllables," in *Proc. Eurospeech-97*, Rhodes, Greece, Sept 22-25. 1997, vol. 3, pp. 1171-1174.

[9] A. Sethy and S. Narayanan, "Split-lexicon based hierarchical recognition of speech using syllable and word level acoustic units," in *Proc. ICASSP-2003*, Hong Kong, Apr 6-10. 2003, vol. 1, pp. 772-776.

[10] A. Sethy, B. Ramabhadran, and S. Narayanan, "Improvements in ASR for the MALACH project using syllable-centric models," in *Proc. IEEE ASRU-2003*, St. Thomas, US Virgin Islands, Nov 30 – Dec 4, 2003.

[11] B. Ruske and G. Plannerer, "Recognition of demisyllable based units using semicontinuous hidden Markov models," in *Proc. ICASSP-92*, San Francisco, California, USA, March 23-26. 1992, I-581 – I-584.

[12] R. Messina and D. Jouvet, "Context dependent "long units" for speech recognition," in *Proc. ICSLP-2004*, Jeju Island, Korea , Oct 4-8. 2004, pp. 645-648.

[13] *TIMIT Acoustic-Phonetic Continuous Speech Corpus*. National Institute of Standards and Technology Speech Disc 1-1.1, NTIS Order No. PB91-505065, 1990.

[14] N. Oostdijk, "The Spoken Dutch Corpus. Outline and first evaluation," in *Proc. LREC-2000*, Athens, Greece. May 31 – 2 Jun. 2000, pp. 887-894.

[15] S. Young, G. Evermann, T. Hain, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, *The HTK Book (for HTK Version 3.2.1)*. Cambridge University Engineering Department, Cambridge, UK, 2002.