The Keyword Lexicon - An accent-independent lexicon for automatic speech recognition

Christophe Van Bael[*] and **Simon King**[†]

[*] A^2RT, Department of Language and Speech, University of Nijmegen, The Netherlands
C.v.Bael@let.kun.nl

[†] Centre for Speech Technology Research (CSTR), University of Edinburgh, UK
Simon.King@ed.ac.uk

ABSTRACT

Recent work at the Centre for Speech Technology Research (CSTR) at the University of Edinburgh has developed an accent-independent lexicon for speech synthesis (the Unisyn project). The main purpose of this lexicon is to avoid the problems and cost of writing a new lexicon for every new accent needed for synthesis. Only recently [1], a first attempt has been made to use the Keyword Lexicon for automatic speech recognition.

1 INTRODUCTION

Modelling pronunciation variation is a hot topic in speech recognition. Previous attempts to handle pronunciation variation at the lexical level, such as [2, 3], largely focused on adding multiple pronunciations per lexeme in the lexica to fit the acoustic data better. Even though it has been proven that this procedure can improve recognition accuracy, it might also lead to a higher degree of lexical confusability, and hence to higher Word Error Rates (WERs).

Assuming that speech characteristics are (at least approximately) known for the training data, the test data or both, the confusability problem may be partly solved by generating accent- or speaker-specific lexica. Our approach exploits the fact that there is at least *some* invariability in a speaker's accent or in several speakers' accents. For example, when a person starts talking about /b A: T/ (SAMPA-notation is used throughout this article), the chances are high that person will prefer the pronunciation /p A: T/ over /p { T/. This consistency may be exhibited within one speaker or a group of speakers such as those from a particular town, region or country. As accent-specific lexica covering these consistencies can be easily built with the Keyword Lexicon, the main aim of this research was investigating its abilities to model pronunciation variation for ASR.

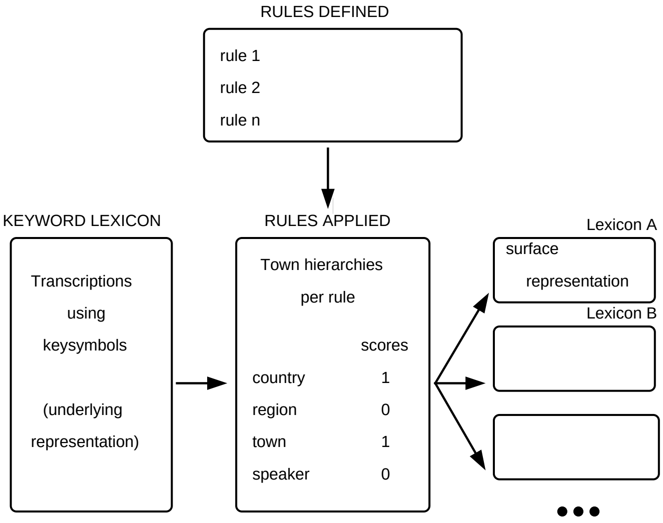

Figure 1: The Keyword Lexicon - Survey

1.1 The Keyword Lexicon

The Keyword Lexicon [4, 5] (figure 1) is an abstract lexicon consisting of keyvowels and keyconsonants following the original keyvowel idea of Wells [6]. Each keysymbol defines a phoneme in a class of words pronounced in a similar way within any accent of English. Hence, the keyvowel *A* defines the vowel in a group of words containing both *path* and *bath*. In the example given, this means that in Leeds *path* and *bath* will both be pronounced with the short vowel /{/, hence /p { T/ and /b { T/. In a typical Southern English accent however, those words will both be pronounced with the long vowel /A:/, hence /p A: T/ and /b A: T/.

Accent-specific lexica can be built by applying hierarchically ordered rewrite rules to the underlying Keyword Lexicon. Figure 2 shows how different lexica can be built by defining whether the rewrite rules apply at the country, region, town and person level. If a rule applies (score 1 in the figure) at a higher level, that rule will be applied when building lexica for that and all underlying levels. Rule scores from higher levels can be explicitly overriden, though, at lower levels.

Hence, in figure 2, the h-drop rule, stating that in some phonetic contexts in some accents of English an initial /h/ may not be pronounced, is not applied to the underlying Keyword Lexicon when building a Northern English (N_ENG) lexicon, while it does apply for all speakers in Newcastle (Newc.), except for speaker Newc.1.

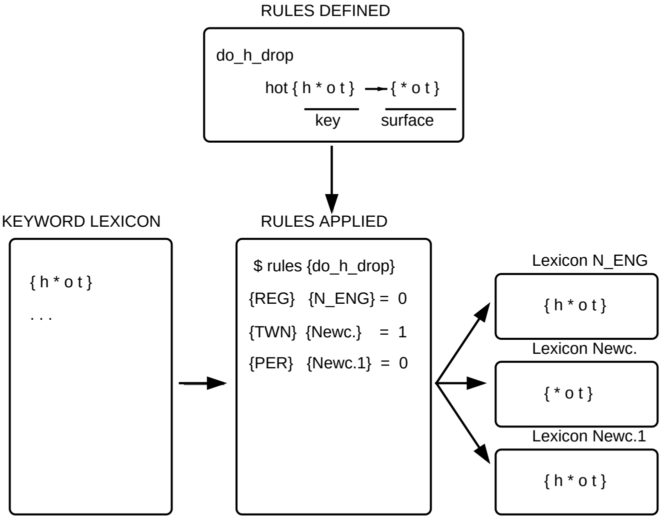

Figure 2: The Keyword Lexicon - Method

2 MATERIAL AND METHOD

2.1 Material

WSJCAM0 [7], the British English version of part of the Wall Street Journal (WSJ) corpus, was used for training, tuning and testing. The training set comprised 7860 utterances, the tuning set (to determine the optimal *language model scaling factor* to scale the influence of the language model and that of the phone models with regard to each other, the optimal *word insertion log probability* to control insertions and deletions, and the optimal pruning) 200 utterances, and the test set 1090 utterances, altogether speech of 140 subjects speaking with various British English accents and defining a 5K closed vocabulary task.

The recogniser was built using Hidden Markov Models (HMMs) with HTK [8]. Word internal tied-state context-dependent models with multiple Gaussian mixture components per state were used. A standard backed-off bigram language model was used at recognition time.

The speech files were sampled at 16kHz and 16 bits per sample. They were parameterised as MFCCs with 12 coefficients to capture the spectral properties of the waveform, one coefficient to capture the amplitude, 13 delta and 13 acceleration coefficients. A standard left-right 3-state topology was maintained for all phone models and a separate silence model, except for a 1-state left-right short pause model used to model the optional silence between words. As no bootstrap data were available, all phone models were built from scratch.

The British English Example Pronunciation (BEEP) dictionary was used. It is a typical multiple pronunciation dictionary of the type usually used for ASR. Out Of Vocabulary words (OOVs) were inserted from the US English Carnegie Mellon University (CMU) multiple pronunciation dictionary after a suitable phone set mapping. The use of the CMU dictionary in this context was clearly a suboptimal solution as the CMU lexicon is an American English lexicon and as BEEP and the WSJCAM0 task cover British English. Due to time restrictions, though, this proved to be the easiest way for us to cover the OOVs.

2.2 Method

The focus of this study is the lexicon. In all experiments, the lexica and the phonetic transcriptions derived from them were the only variables. This means that the data, the language model and the recognition task as a whole remained the same at all times.

To test whether the Keyword Lexicon was suitable for recognition, its performance was compared to a baseline. This baseline represents the *traditional* approach of modelling pronunciation variation in ASR at the lexical level: the use of lexica with multiple pronunciations per lexical entry (multiple pronunciation lexica). This baseline was obtained using the BEEP lexicon. The same recognition task was then performed with a conglomerate of all 8 pre-defined English lexica covered by the rule-sets in the first release of the Keyword Lexicon. This will be called the *merged keyword lexicon* hereafter. In this way the Keyword Lexicon was tested as a multiple pronunciation lexicon, thus neglecting its possible gain for pronunciation variation modelling (its ability to easily create accent-specific lexica) for the time being.

After testing the Keyword Lexicon's inherent suitability for ASR by using it as a multiple pronunciation lexicon, its true contribution to ASR was investigated in 2 experiments using the Lexicon as an abstract lexicon from which accent- or speaker-specific lexica were generated. In the previous experiments the same lexicon was used both for training (to generate phone transcriptions from the training data through a lookup procedure) and for recognition (as a top-down constraint). However, the lexica used for training and recognition may also differ per experiment. In that way more specific lexica can be used that better fit the train and test data. We adopted this procedure in the following experiments.

In the first experiment testing the Keyword Lexicon's quality and ease of use, the merged keyword lexicon was used for training, while at recognition time accent-specific lexica generated from the Keyword Lexicon were used. In the second experiment the training speakers were roughly divided into 7 accent groups, and for each one of these groups, an accent-specific lexicon was generated. These lexica were used to train the acoustic models with, while the same accent-specific lexica of the previous experiment were used to test the recogniser.

3 BASELINE AND PRELIMINARY EXPERIMENT

3.1 Baseline: recognition with the BEEP lexicon

1512 words of the training corpus and 66 words of the test corpus did not appear in BEEP. Therefore, the (multiple) pronunciations of these words were inserted from the CMU lexicon. The total phone set comprised 46 phones. The least frequent phone occured 57 times in the training lexicon, the most frequent phone 5583 times. The first baseline was 31.4% WER.

3.2 Testing the Keyword Lexicon as a multiple pronunciation lexicon

The merged keyword lexicon (see 2.2) was used both for training and testing the recogniser. Again, the missing words were inserted from the CMU dictionary. The total phone set, however, now comprised 83 phones. This is a large set, given that the same amount of training data was available as for the baseline experiment (with the BEEP lexicon and only 46 phones). All phones were needed, though, to capture the most important pronunciation variants in the merged keyword lexicon. After conducting the same recognition task a WER of 32.5% was obtained. A two-tailed t-test proved that this WER was not significantly different from the 31.4% obtained when testing with the BEEP lexicon.

4 TESTING THE KEYWORD LEXICON

For these experiments, 7 accent-specific lexica were built from the Keyword Lexicon to cover the training data, 4 accent-specific lexica to cover the development test set and 7 more accent-specific lexica to cover the speakers from the test set not covered by any of the other lexica. The accents covered are presented in table 1. Whereas the division of the training speakers in accent-groups was a coarse one based on geographical information of the speakers, the division of the test speakers into accent groups was based on the judgements of two expert listeners.

Train set	Tuning set	Test set
Southern, Northern, Central 1, Western, Welsh, Irish, Scottish	Southern, Northern, Central 1, RP 1	RP 1,2,3,4 Central 1,2,3,4 Western, Irish, Scottish (Edi)

Table 1: Accent groups in the data sets.

4.1 Training with a multiple pronunciation lexicon, testing with accent-specific lexica

Training was performed with the merged keyword lexicon, as in 3.2. For recognition, the test speakers were divided in 11 accent-groups, and per accent-group an accent-specific lexicon was generated. With a WER of 39.7%, in this experiment the use of the Keyword Lexicon led to a decrease in accuracy compared to the same task performed with multiple pronunciation lexica.

4.2 Training and testing with accent-specific lexica

For this experiment, the training data were subdivided in 7 accent groups, and new acoustic models were trained from phonetic transcriptions generated with 7 accent-specific lexica. However, as the development test set only comprised 3 different accents whereas the test set comprised 11 different accents, and as the recogniser had to be tuned again because new phone models were used, in this and in the following experiment the recogniser was tuned on the test set. Therefore we are careful interpreting the results, but we believe that the 32.5% WER obtained in this experiment at least indicates that the use of the Keyword Lexicon can lead to equally good recognition results as the use of multiple pronunciation lexica.

5 DISCUSSION

In the first experiment BEEP was used both for training and for recognition. Figure 3 illustrates the recogniser's baseline performance (31.4% WER) on the separate accent-groups.

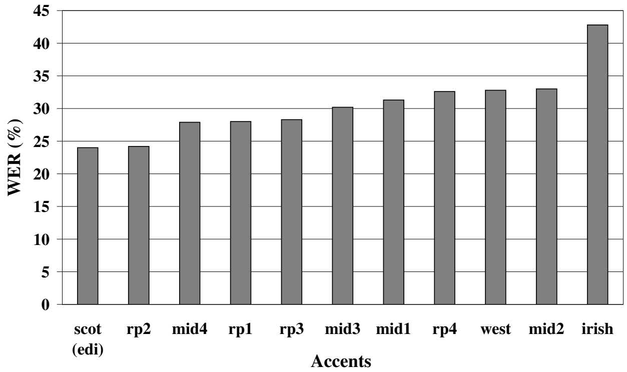

Figure 3: Baseline: accent-specific results

The second experiment tested the Keyword lexicon as a multiple pronunciation lexicon. The performance (32.5% WER) was not significantly different from the baseline. When comparing the results of the baseline experiment and the results of this experiment per accent, the WERs resembled. Therefore we conclude that the Keyword Lexicon is a worthy equivalent of BEEP when used as a multiple pronunciation lexicon for ASR.

Using a multiple pronunciation lexicon for training and accent-specific lexica for recognition resulted in a significantly worse performance (39.7% WER) than the baseline. Again a similar distribution was discovered when comparing this performance with the baseline at a speaker- and accent level. The results could mean three things: 1) the pooled training lexicon wasn't suitable to train phone models (either it didn't fit the data or the degree of confusability between the different pronunciation variants was too high), 2) the accent-specific test lexica didn't fit the data or 3) our new approach to ASR proposed here had to be reconsidered.

Finally, accent-specific lexica were used both for training and for recognition. Using accent-specific training lexica does not involve any extra effort in our approach, as those lexica can be made swiftly based on information of the training speakers. Now a WER of 32.5% was obtained. Putting more time and effort in determining the training speakers'accents and preferably even their personal characteristics and incorporating this knowledge into speaker-specific rule sets to generate more specific training lexica, might further improve the recogniser's performance, thus outperforming multiple pronunciation lexica on the same task. Even though one has to be careful interpreting these results (in this experiment the recogniser was tuned on the test set), we conclude that the problem in the previous experiment was the multiple pronunciation training lexicon, and not the accent-specific test lexica, nor our new approach as such. This experiment proves that our new approach can at least compete with the use of multiple pronunciation lexica on the same task.

6 CONCLUSION

We have introduced a new approach to model pronunciation variation for ASR at the lexical level. Until now, research has largely focussed on creating large multiple pronunciation lexica for speech recognition. However, when too many pronunciation variants are included, the confusability increases, and the recogniser's performance decreases. Our new approach introduces an abstract lexicon from which accent- or even speaker-specific training and test lexica can be generated via rule-settings that can be defined with great ease.

The experiments showed that the Keyword Lexicon serving as a multiple pronunciation lexicon performs equally well as a standard multiple pronunciation lexicon like the BEEP dictionary. Moreover, using still quite general accent-specific lexica both for training and for recognition gave a similar performance on the same task. We believe that adding more time and effort in building speaker-specific training lexica might easily improve the recognition accuracy, without assuming more knowledge of the test speaker's speech characteristics.

ACKNOWLEDGEMENTS

The research reported was performed as part of a Masters program between May and September 2002 at the Department of Theoretical and Applied Linguistics at the University of Edinburgh, UK.

REFERENCES

[1] C.P.J. Van Bael, "Using the Keyword Lexicon for speech recognition," M.Sc. thesis, The Department of Theoretical and Applied Linguistics, University of Edinburgh, 2002, available at: http://lands.let.kun.nl/literature/vanbael.2002.2.pdf.

[2] J.M. Kessens, *Making a difference. On automatic transcription and modeling of Dutch pronunciation variation for automatic speech recognition*, Ph.D. thesis, University of Nijmegen, The Netherlands, 2002.

[3] M. Wester, *Pronunciation variation modeling for Dutch Automatic Speech Recognition*, Ph.D. thesis, University of Nijmegen, The Netherlands, 2002.

[4] S. Fitt and S. Isard, "Synthesis of regional English using a keyword lexicon," in *Eurospeech 99*, 1999.

[5] S. Fitt, "Documentation and user guide to UNISYN lexicon and post-lexical rules," Tech. Rep., Centre for Speech Technology Research, University of Edinburgh, 2000.

[6] J.C. Wells, *Accents of English*, Cambridge University Press, Cambridge, 1982.

[7] J. Fransen et al., "WSJCAM0 corpus and description," Tech. Rep. CUED/F-INFENG/TR.192, Cambridge University Engineering Department (CUED) Speech Group, Trumpington Street Cambridge CB2 1PZ U.K., 1994.

[8] S. Young et al., "The HTK book (for HTK version 3.1)," Tech. Rep., Cambridge University Engineering Department, 2001.