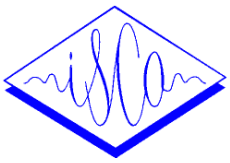

6th International Conference on Spoken Language Processing (ICSLP 2000)
Beijing, China
October 16-20, 2000

IMPROVED LEXICON FORMATION THROUGH REMOVAL OF CO-ARTICULATION AND ACOUSTIC RECOGNITION ERRORS

Philip Hanna, Darryl Stewart, Ji Ming & F.J. Smith

School of Computer Science
The Queen's University of Belfast, Northern Ireland

ABSTRACT

It is becoming increasingly more necessary that speech recognition systems contain an accurate lexicon, consisting of likely word pronunciations that actually occur within a given domain. Given the increasing size of speech databases, it would appear that data driven approaches are best suited to derive such pronunciations. Presently, however, such an approach often introduces implausible pronunciations, resulting in a higher degree of confusability within the decoder. In this paper, we outline a novel data driven approach which aims to improve the quality of extracted word pronunciations through the removal of co-articulation effects and acoustic model misclassifications from the speech data. A number of selection constraints are additionally employed to exclude any improbable pronunciation alternatives. Initial experiments have shown that the approach does indeed provide plausible pronunciation alternatives without introducing improbable pronunciations.

1. INTRODUCTION

Undeniably the lexicon is one of the main sources of knowledge within a speech recognizer, but yet it is often neglected. Arguably, a lexicon solely consisting of canonical pronunciations may have proved sufficient for early speech recognition systems, however, for present day systems dealing with spontaneous or conversational speech this is not the case.

Spontaneous speech effects have been well documented, e.g. [1]. Such effects entail that certain words are observed to have a varied range of pronunciations. Evidently, the accuracy with which the lexicon models different word pronunciations has a direct bearing upon the overall effectiveness of the decoder.

Initially, expert driven approaches were used to improve the lexicon quality, whereby a linguist would augment the lexicon with alternative pronunciations extracted from a body of speech. However, as the quantity of speech data employed by speech recognition systems substantially increased, the use of expert driven approaches became less feasible, leading to the introduction of data driven approaches. Typically, data driven approaches derive pronunciations by mapping expected word sequences onto recognized phone sequences (i.e. an expected phone sequence is built up from a concatenation of canonical pronunciations of the expected word sequence, which is then mapped onto the recognized phone sequence via a DP process). All observed pronunciations of a given word are then collected, and those pronunciations that occur with sufficient frequency are added to the lexicon. Typical examples can be found in [2,3].

Whilst data driven approaches can be readily applied to large speech databases the quality of the alternative pronunciations produced by such systems may be relatively poor when compared to those determined by a linguist [4]. There is also a tendency that either too few or too many pronunciation alternatives are proposed [4], providing the decoder with sub-optimal search constraints.

The goal of this paper is to outline a novel data driven approach that aims to derive accurate pronunciation alternatives without consequentially introducing any implausible pronunciations. Linguistically, pronunciations differing from the canonical can be ascribed to either intra- or inter-word variations (see Figure 1). Inter-word variations are due to effects such as co-articulation, whereby the surrounding phone context can lead to phone insertions, deletions or substitutions. Intra-word variations arise as people with different regional accents, etc., tend to pronounce certain words differently. Importantly, a third source of variation arises from acoustic model misclassification, i.e. an HMM may delete, insert or substitute phones. Although acoustic misclassification has no bearing on how words are actually pronounced, it may nevertheless predictably mutate the recognized phone sequences (e.g. the acoustic models may readily confused the /n/ and /m/ phones).

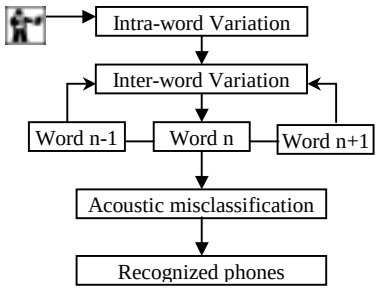

Figure 1. Simple block diagram of inter- and intra-word pronunciation variation and acoustic misclassification.

In [5], Sloboda outlines a system which employs an HMM phone confusion matrix in order to remove alternative pronunciations which can be ascribed to simple acoustic misclassification of a canonical pronunciation. We describe an approach that can be viewed as an extension of this idea, whereby co-articulation effects and acoustic misclassifications are probabilistically removed from the observed phone transcripts in order to provide pronunciations solely due to intra-word variations. Furthermore, through the application of various selection constraints it is possible to reduce the number of implausible pronunciations.

The remainder of this paper is organized as follows: In section 2 brief details are given of earlier work concerning the modeling of co-articulation effects and acoustic misclassifications. In section 3 an outline is given of how pronunciation alternatives may be extracted, both in terms of the removal of co-articulation and acoustic misclassification and also pronunciation selection constraints. Finally, in section 4 some experimental results are presented.

2. MODELLING OF CO-ARTICULATION EFFECTS AND ACOUSTIC MISCLASSIFICATIONS

Central to the approach outlined in this paper is the ability to derive statistical information concerning co-articulation effects and acoustic misclassification. Such statistics are valuable in themselves, and can be readily applied within the decoder to take into account the various ways co-articulation may effect a word pronunciation or the likely errors which the acoustic model may introduce. Given that co-articulation and acoustic misclassification tend not to be word dependent, but rather at best dependent upon the surrounding phone context, it is desirable that they be modeled separately from lexical word pronunciations.

Previous research by the authors, e.g. [6], has provided a data driven method through which co-articulation and acoustic misclassification can be modeled, i.e. individual phone mutation (insertion, deletion or substitution) probabilities dependent upon a surrounding phone context are derived. For example, the probability of /ae/ being deleted given a left and right phone context of /d/ and /er/ respectively can be determined. Note that the probabilities of correctly identifying phones are also determined, i.e. match probabilities.

The phone mutation probabilities are collected through an iterative procedure. On the first iteration a flat set of probabilities are assumed (e.g. all deletions are given the same starting probability). Each iteration uses a DP mapping procedure in conjunction with the mutation probabilities to map recognized phone transcripts onto expected phone transcripts. At the end of each iteration the mutation probabilities are modified, depending upon how frequently they were employed, as follows:

$$\text{Probability}_{\text{m}} = \frac{f_m}{f_c} \times (1 - e^{-\alpha f_c}) \quad (1)$$

where f_m is the frequency with which the phone mutation occurred for a given surrounding phone context, f_c the total number of times the context was observed irrespective of any insertions, deletions or substitutions, and α is a weighting term used to lower the probabilities when the context occurs infrequently. A set of baseline probabilities are used under which no mutation probability can fall, catering for those cases where no instances were observed.

The outlined iterative procedure is repeated until the probabilities converge or a set number of iterations have passed.

3. GENERATION OF PRONUNCIATION ALTERNATIVES

The proposed method of generating alternative pronunciations can be separated into two processes. Firstly, removal of co-articulation and acoustic misclassification effects. Secondly, application of additional selection constraints in order to remove any improbable pronunciations. In practice, both processes are employed in conjunction with one another.

3.1 Removal of Co-articulation and Misclassification Effects

The generation of alternative pronunciations can be separated into three stages, as shown in Figure 2. The process assumes the availability of a base lexicon, which is to be augmented with alternative pronunciations drawn from a quantity of acoustically transcribed speech (i.e. as recognized by an HMM).

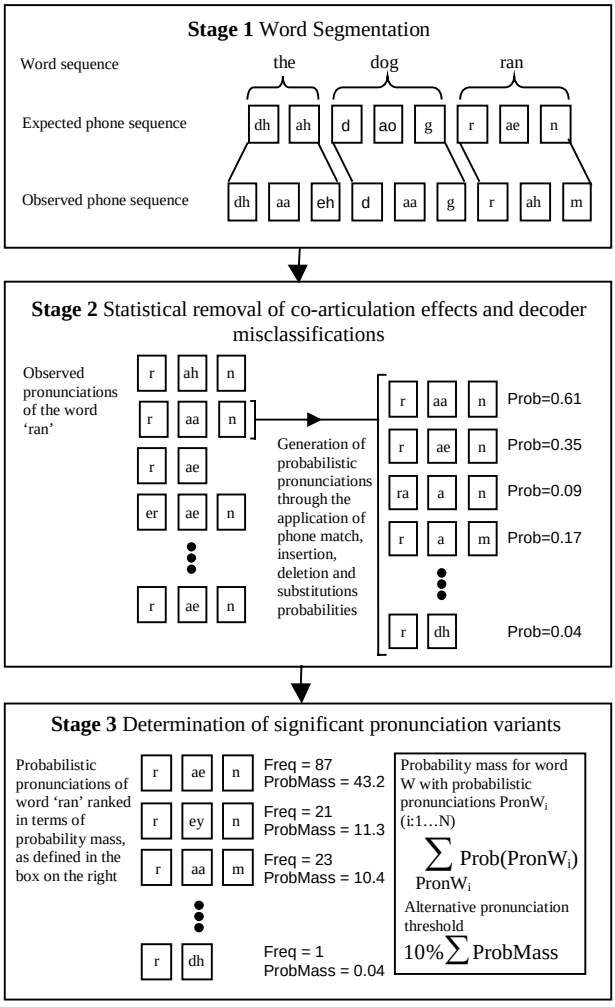

Figure 2. The three proposed stages of alternative pronunciation generation.

A breakdown of the stages shown in Figure 2 follows:

Stage 1. The first stage involves segmenting the observed phone sequence into phone strings corresponding to word pronunciations. This is accomplished through the application of a DP mapping between the observed phone transcript and the expected phone sequence built up from a concatenation of pronunciations as predicted by the base lexicon. In the case where several expected phone sequences can be formed, i.e. the base lexicon contains multiple pronunciations of a word, then the expected sequence that produces the best DP match with the observed phone sequence is employed.

Stage 2. The second stage aims to probabilistically remove the effects of co-articulation and HMM misclassifications. The process is probabilistic as it uses the phone mutation statistics (as outlined in section 2) to construct probabilistic pronunciations of an observed word pronunciation. Hence, for a given utterance of a word, a further list of pronunciations is generated representing possible utterances which were then subjected to co-articulation effects and acoustic misclassifications. For example, an observed pronunciation of the word 'ran' might be /r/ /aa/ /n/. Two probabilistic pronunciations of this word are /r/ /ae/ n/, representing the case where /aa/ has been substituted for /ae/, and /r/ /aa/ /n/ where the utterance has been correctly recognized and not subject to co-articulation.

It is necessary to impose a number of constraints throughout this stage in order to restrict the number of probabilistic pronunciations which are generated. In particular, any phone mutation probabilities which fall under a predefined minimum threshold are not applied. Furthermore, it is assumed that a given word pronunciation cannot be subject to more than a certain maximum number of mutations (the maximum is determined as a constant value plus a variable amount dependent upon the number of phones in the pronunciation).

The probabilistic expansion is applied to all pronunciations of a given word, over all words in the lexicon for which a certain minimum number of pronunciations of that word can be collected (in order to ensure statistically meaningful results). At the end of this stage, the probabilistic pronunciations of each word are accumulated together.

Stage 3. The final stage aims to select, for a given word, those pronunciation alternatives which are significant from the accumulated lists of probabilistic pronunciations. In order to do this it is not only necessary to consider the number of occurrences of a possible alternative pronunciation but also the probabilities of those occurrences, i.e. a pronunciation with a high probability is considered more significant than a pronunciation with a lower probability.

As shown in Figure 2, the probability mass of word W with probabilistic pronunciations $\text{Pron}W_i$ (i:1…N), is determined as follows:

$$\sum_{\text{Pron}W_i} \text{Prob}(\text{Pron}W_i) \qquad (2)$$

Any pronunciations of a word, whose probability mass accounts for more than a predefined percentage of the summed probability mass of all pronunciations of that word, is considered to be a significant pronunciation. This can be further constrained by requiring that significant pronunciations must occur a certain minimum number of times. The probability mass associated with a pronunciation provides a means of estimating the likelihood of that variant.

The above process permits significant word pronunciations to be extracted. There are a number of different options available as to how the significant pronunciations are combined with the starting lexicon. In principle, the outlined approach could be used to generate an entirely new lexicon, if given sufficient data. In practice, any alternative pronunciations classified as significant will be added to the existing base lexicon.

3.2 Constraining the Selection of Significant Pronunciation Alternatives.

The process as outlined above will remove co-articulation effects and acoustic misclassifications from the observed pronunciations. However, an unwelcome side effect is that pronunciations not subjected to such phenomena may have co-articulation effects or acoustic misclassifications added. This can lead to the introduction of spurious, poor pronunciation alternatives.

For example, a trained speaker may clearly pronounce the word 'and' as predicted by the base lexicon, e.g. /ae/ /n/ /d/. Should the phone mutation statistics give a high probability that /n/ and /m/ can be confused, then a likely, but unwanted alternative pronunciation will be /ae/ /m/ /d/. Such behavior can be largely removed through the use of additional selection constraints, as follows:

- Any observed pronunciation of a word that exactly matches an existing base lexicon pronunciation should not undergo the expansion outlined in stage 2. Presumably such pronunciations have no co-articulation, etc. to remove.
- Any alternative pronunciations produced at the end of stage 3 which are sufficiently similar (as defined below) to existing base lexicon pronunciations are considered to be simply an utterance of the base lexicon item subjected to co-articulation effects or acoustic misclassification. As such, they are discarded, and not added to the lexicon. As an aside, the use of the phone mutation probabilities during decoding will ensure that such pronunciations are likely to be mapped onto the correct lexical pronunciation.
- Any two alternative pronunciations of a given word which are sufficiently similar (as defined below) will be collapsed to the most significant of the two pronunciations. The combination can be made in terms of which pronunciation has the largest probability mass, or alternatively, the largest number of occurrences.

For the purposes of constraining the pronunciation alternatives, a DP mapping can be used to determine if one pronunciation might simply arise from co-articulation variation or acoustic misclassification of another pronunciation. In particular, one pronunciation is considered similar to another if more than a certain percentage of the DP phone mappings have a probability over a predefined threshold.

4. RESULTS

The difficulties associated with lexical assessment techniques, mostly in terms of comparability and applicability, have been soundly discussed elsewhere [7]. The problems are understandable; in order to test the effectiveness of knowledge relating to likely word pronunciation it is necessary to apply that knowledge to a particular recognition task. Evidently, the efficiency with which the knowledge is applied, coupled with the difficulty of the recognition task have a direct bearing upon the overall reported performance.

The results presented here are measured in terms of metrics that attempt to more clearly divorce the effects of the decoding system from measured lexical performance. In particular, two metrics are employed; the first measures the performance (in %word accuracy) of mapping recognized phone sequences onto the corresponding expected phone sequences; the second measures the increase in lexical confusability (in terms of how many valid word hypothesis, other than the expected, are introduced). The metrics can be more precisely defined as follows:

%word mapping accuracy: Given a quantity of recognized phonetic transcripts and a set of expected word sequences, a segmentation is firstly performed to segment the phonetic sentence transcripts into recognized word pronunciations.

If a segmented word pronunciation correctly maps onto any lexical pronunciation of the expected word, then a successful mapping is recorded. However, the phone mutation statistics can also be applied to predict likely-co-articulation or acoustic misclassifications. For a given phone sequence, the *n* most likely phonetic interpretations of that phone sequence can be derived. Should any of the *n*-best sequences map onto an expected lexical entry, then a correct mapping is recorded.

By varying *n* it is possible to simulate decoders of different search complexity. Evidently, the more accurate the phone mutation statistics, the lower the value *n* must assume.

Lexical confusability: Lexical confusion can likewise be measured given the above framework. In particular, for an *n*-best list of expanded phone sequences, if any phone sequences map onto lexical entries other than the expected - and have a higher rank than the expected mapping - then they are considered as lexical confusions, i.e. a valid word hypothesis that must be excluded within the decoder using other knowledge sources.

Based on the use of the above two metrics, results have been collected for the WSJ and Timit databases. In both cases mono-phone acoustic models were employed, thereby ensuring a considerable degree of phonetic variation within the transcripts (52% phone recognition accuracy for WSJ and 56% for Timit).

The baseline system consisted of a canonical lexicon. This was then aided through the use of phone mutation statistics and finally augmented with new lexical pronunciations. Results for the WSJ and Timit databases follow (an *n* value of 25 was used throughout all experiments).

Table 1 – Lexical performance and confusability for WSJ and Timit databases, $n = 25$

Database/Model		%Word Mapping Accuracy	Increase in lexical confusability
Timit	Baseline	19	--
	Baseline + Phone mutation statistics	38	184%
	Augmented lexicon + Phone mutation statistics	47	263%
WSJ	Baseline	23	--
	Baseline + Phone mutation statistics	41	170%
	Augmented Lexicon + Phone mutation statistics	55	212%

5. SUMMARY

In conclusion, the experimental findings have shown that the phone mutation probabilities (modeling co-articulation effects and acoustic misclassifications), when used in conjunction with the alternative pronunciations, provide a capable means of determining how a word is likely to be pronounced. This increase is at the expense of more lexical confusability, however, additional analysis of the results has shown that a language model can exclude lexically confusable items in the majority of cases.

The authors are presently extending this research towards the use syllabic data. For example, the outlined alternative pronunciation procedure may be improved by introducing syllable level constraints (with regards to the location of co-articulation errors within the syllable, word-rule formation, etc.).

6. REFERENCES

[1] Greenberg, S. “Speaking in Shorthand – A Syllable-Centric Perspective for Understanding Pronunciation Variation”. *Proceedings of the ESCA Workshop on Modeling Pronunciation Variation for Automatic Speech Recognition*, Kekrade, pp.47-56, 1998

[2] Fosler, E., Weintraub, M., Wegmann, S., Kao, Y-H., Khudanpur, S., Gallos, G. & Saradar, M. “Automatic Learning of Word Pronunciations from Data”. *Proceedings of ICSLP’96*, SaP2S1.2, 1996

[3] Ravishankar, M. & Eskenazi, M. “Automatic Generation of Context-Dependent Pronunciations”. *Proceedings of Eurospeech’97*, pp.2467-2471, 1997

[4] Byrne, W., Finke, M., Khudanpur, S., McDonough, J., Nock, H., Riley, M., Saraclar, M., Wooters, C. & Zavaliagkos, G. “Pronunciation Modeling Using a Hand-Labeled Corpus for Conversational Speech Recognition”. *Proceedings of ICASSP’98*, pp.I313-316, 1998

[5] Sloboda, T. “Dictionary Learning: Performance Through Consistency” *Proceedings of ICASSP’95*, pp.453-56, 1995

[6] Hanna, P., Stewart, D., Ming, J. & Smith, F.J. “An Improved DP Match for Automatic Lexicon Generation”, *Proceedings of ICPhS 99*, pp.1717-1720, 1999

[7] Strik, H.,Kessens, J.M., Wester, M. “Modeling Pronunciation Variation for Automatic Speech Recognition” *Proceedings of the ESCA Workshop*, Kerkrade, pp.137-144, 1998