An Improved DP Match for Automatic Lexicon Generation

Philip Hanna, Darryl Stewart, Ji Ming, F. Jack Smith
The Queen's University of Belfast, UK

ABSTRACT

A number of automatic lexicon construction methods have been proposed in recent years. Such approaches employ a dynamic programming (DP) match to collect statistics concerning differences between the observed phone sequence and that which was predicted by a standard lexicon. A more expressive lexicon is then constructed based upon the collected statistics for use within a speech recognition system, which offers a more accurate phone-to-word mapping. We show that the standard DP procedure leads to the introduction of spurious matches, which reduces the quality of any subsequent processing based upon the DP provided matches. In order to remove this deficiency, an iterative DP match procedure, using individual phone confusion probabilities is outlined. It was found that the iterative DP match significantly reduced the number of equi-probable matches, to the extent that for the vast majority of utterances, only one possible DP mapping resulted, thereby improving the quality of generated statistics.

1. INTRODUCTION

In recent years a number of researchers have turned their attention towards methods of improving the quality of the lexicons that are used within speech recognition systems. Typically, this is achieved by augmenting a standard lexicon with common pronunciation variants and/or phonological rules [1,2,3]. These lexical extensions are increasingly more necessary given the complexity of modern day, speech recognition tasks. It has been estimated that for spontaneous speech approximately 40% of all words are not pronounced as would be predicted by a standard lexicon [4] that implicitly assumes each word is pronounced under good acoustic conditions, by a trained speaker. Evidently, if the lexicon more accurately models how words are expressed in terms of more simple acoustic units, then improved recognition performance should result.

A loose trend may be observed in the development of lexicon construction techniques; commencing with expert driven approaches (e.g. [5]) which made use of linguistics to select pronunciation variants and/or phonological rules, through to semi-automatic approaches (e.g. [6]) which employ an automatic method to suggest improvements, from which a linguist would select the most promising. Finally, fully automatic approaches aim to remove the need for expert involvement, whilst still returning the same quality of results as would a linguist. This is typically accomplished through the use of a DP algorithm to match a number of observed phone sequences onto their corresponding expected phone sequences (as predicted by a standard lexicon). Differences between the expected and observed phone sequences, expressed in terms of phone deletions, insertions and substitutions, are analyzed in order to model co-articulation effects, whereas a study of all the observed pronunciations of a word, leads to the extraction of alternative pronunciations of that word.

Fully automatic approaches are desirable as they can be easily applied to new and large databases, for which it is impractical to use an expert driven approach due to the necessary time involvement, and inconsistencies arising as different linguists are unlikely to produce exactly the same classification [7]. Semi-automatic approaches also suffer from the same limitations, albeit to a lesser degree. However, whilst fully automatic approaches are attractive, it is widely recognized that they tend to only provide marginal performance increases [8,9]. In this paper, we identify one source attributing to their lackluster performance, and show how it may be overcome.

This paper is organized as follows: In section 2, we explore the limitations of the present automatic approaches, with emphasis on the DP match. Section 3 outlines the improved DP technique, the results of which are presented in Section 4. Finally, we briefly summarize our findings.

2. LIMITATIONS OF PRESENT AUTOMATIC APPROACHES

Researchers have identified various problems with the present automatic lexicon formation techniques, and also how they may be overcome [7,8]. The problem most often highlighted is that of too few, or alternatively too many, rules/pronunciations, resulting in either inadequate coverage of possibilities, or overgeneralization which increases the amount of confusability. It has been suggested [8] that introducing constraints, for example associating a weighted penalty with each alternative pronunciation or rule, would help to more accurately constrain the search. Furthermore, the HMM misclassifications also significantly contribute to the amount of observed phone insertions, deletions and substitutions, and should be accounted for.

To this end, optimal results would be expected to arise from an approach which utilizes a large number of phone transcriptions generated from the same set of acoustic models as will eventually be employed during recognition, and additionally assigning weighted penalties to not only co-articulation effects, but also HMM misclassifications. However, whilst such a system is capable of capturing and discriminatively employing a wide range of information sources, i.e. offering a better application of the lexical information, it does not consider the quality of the statistics collected through the DP match. To this end, we identify a problem with the standard application of the DP algorithm, and show how it may

be overcome, thereby providing a means of collecting more statistically accurate information.

2.1 Limitation of the DP algorithm

The DP problem arises, as typically within the DP match phone insertions, deletions and substitutions are each assigned a pre-selected penalty. This is problematic, and often results in spurious matches between an expected and observed phone sequence, as illustrated in Figure 1, which assumes deletions, insertions and substitutions are equally penalized. The application of a DP match to generate an optimal mapping of the observed phone sequence onto the expected phone sequence resulted in a total of three equi-probable mappings. Some of the matches might be undesirable if subjected to further scrutiny (entailing that if a more precise knowledge of the probabilities of phone insertions, deletions and substitutions were available then a smaller subset of possible paths through the network would be obtained). Note, that ∅ represents a null phone.

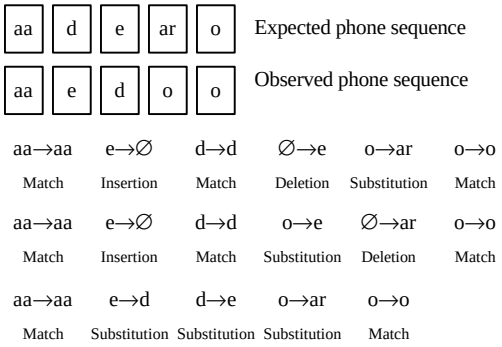

Figure 1 – DP match generation of three equi-probable mappings of the observed phone sequence onto the expected phone sequence, assuming equal insertion, deletion and substitution penalties.

As the DP match is used to generate differences between the observed and expected phone sequences, the effectiveness of any subsequent processing depends upon the quality of the DP produced mappings.

In order to improve the quality of the mappings produced by the DP procedure, we propose that each specific insertion, deletion or substitution for a given phone context, be assigned a probability of occurrence based upon the relative frequency with which it is observed to occur. Such statistics may be obtained through an iterative application of the DP match, whereby a set of insertion, deletion and substitution probabilities are initially assumed, and refined based upon the generated frequency counts. The refined set of probabilities is reapplied within another DP match, either over a fixed number of iterations, or until no significant changes are observed in the probabilities of phone insertions, deletions and substitutions.

3. ITERATIVE DP MATCH TECHNIQUE

The proposed method builds upon the standard DP approach, and as such assumes the existence of a set of acoustic models, so as to permit phone level transcriptions to be generated (i.e. the observed phone sequence) through unconstrained phone recognition. It is also assumed that there exists some form of standard lexicon for the given database, which is to be improved through the automated process, and which also provides the expected phone sequences based on a lexical expansion of the word level transcriptions.

Confusion matrices were constructed detailing the frequency with which each phone was inserted, deleted or substituted for a given phone context. The frequencies were collected using a DP match employing the present set of mutation probabilities. For a given surrounding phone context, the insertion, deletion and substitution probabilities can be expressed as:

Deletion: $[a_i \rightarrow \varnothing \mid a_{i-L},\ldots,a_{i-1}, a_{i+1},\ldots,a_{i+R}]$
Insertion: $[\varnothing \rightarrow a_i \mid a_{i-L},\ldots,a_{i-1}, a_{i+1},\ldots,a_{i+R}]$
Substitution: $[a_i \rightarrow a_j \mid a_{i-L},\ldots,a_{i-1}, a_{i+1},\ldots,a_{i+R}]$

As mentioned, initially all probabilities are assigned a default value corresponding to the type of mutation. However, once frequency counts have been collected each probability is then modified based on its relative frequency of occurrence. Minimum probabilities were imposed to account for those cases where the probability of an uncommon mutation is set to zero if no observations of that mutation were recorded.

Although the use of relative frequencies permits an accurate modification of the probabilities by taking into account the number of times the context appeared without any resultant insertions, deletions or substitutions, problems can nevertheless arise when only limited observations are available. For example, if a particular insertion occurred twice for a given context that was only observed twice, then it would be unwise to assign a probability of one to this insertion given the limited number of occurrences, and hence confidence, upon which the change would be based. In order to overcome this problem a confidence based probability modification was employed, whereby the probability change is dependent upon the number of observations. The confidence based measure is shown in (1), where f_m is the frequency with which the phone change occurred for a given surrounding phone context, and f_c the total number of times the context was observed irrespective of any insertions, deletions or substitutions.

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

As the number of observed contexts increases then so does the confidence with which probabilities can be asserted. If only a limited number of observed contexts are available then the probability estimation will err on the side of caution. Given the above framework, an iterative application may be straightforwardly implemented, whereby the newly modified insertion, deletion and substitution probabilities are reapplied in a subsequent DP match, from which new frequency counts are extracted, and ultimately a new set of probabilities generated. This process may be terminated after a fixed number of iterations or when convergence is obtained.

4. RESULTS & DISCUSSION

This section presents experimental findings for the improved DP technique. The results justify the proposed method and provide a clear insight into the potential capabilities of the approach.

The experiments were conducted in order to determine if the procedure reduces the number of equi-probable DP matches through the refinement of statistical probabilities and also to determine the extent to which the iterative DP match is capable of extracting meaningful statistics.

4.1 Experimental Conditions

All experiments were based on the well-known TIMIT database. This database has been widely used by researchers when exploring means of assessing pronunciation variation [2,8]. This is primarily due to the availability of an expert, hand-labeled, phonetic transcription of the database, which not only permits good quality phone segmentation but also the straightforward extraction of phonological rules. Additionally, the TIMIT database was designed to be phonetically rich, and hence contains a wide variety of co-articulatory effects.

The experiments make use of the hand-labeled transcripts. This entails that the presented results should mostly highlight HMM misclassifications, noise, and any co-articulation effects not fully accounted for by the experts. The observed phone sequences were generated from a set of standard HMM models; each model consisting of a left-to-right, three state topology employing 32 mixtures and using MFCC features augmented with accelerated spectral coefficients. HMM models were constructed for 48 phones during training, which were then folded during recognition, by merging allophones, into the common set of 39 phones.

Initially the probabilities of all insertions, deletions and substitutions were set to 0.1. A minimum probability of occurrence of 0.001 was employed for these mutations, catering for those cases where no occurrences of a particular mutation were observed.

4.2 DP Path Reduction Results

Three different phone contexts were used; one left and one right phone context (1-1), one left phone only context (1-0), and a null phone context (0-0). The particular context that is employed has a strong bearing on the effectiveness of the convergence process. As the context increases in size, so does the number of possible insertions, deletions, substitutions and matches that have to be modeled, and hence the amount of training data necessary to adequately estimate the probabilities. Too little data entails that good probability estimates cannot be formed, resulting in a smaller reduction in the number of equi-probable DP matches.

Figure 2 provides a breakdown of the number of equi-probable DP matches observed across the 624 training sentences. Results are provided for the first 4 iterations (after which little change in the distribution was observed).

As would be expected the distribution of equi-probable DP matches for the 1st iteration is the same across all three contexts; the results being obtained from the same initial probabilities. Furthermore, a brief analysis shows that similar sets of results were obtained for the (0-0) and (1-0) phone contexts, a large reduction in the number of equi-probable paths was observed for the 2nd iteration, followed by smaller changes in path reductions after subsequent iterations, entailing that the process had converged as far as possible given the available training data. The results for (1-1) context show a small change between the 1st and 2nd iterations, with subsequent iterations alternating between two different states. This behavior is most likely due to the large number of phone contexts to be modeled (i.e. there are 39^4 = 2,313,441 possible phone substitutions) coupled with the limited training data.

Number of Paths

0 Left 0 Right Phone Dependencies (0-0)

Iteration No.	1	2	3	4-6	7-10	11-20	21-50	51-100	101-1000	>1000
1	124	104	64	121	56	66	52	22	14	1
2	598	22	1	3	0	0	0	0	0	0
3	592	28	1	3	0	0	0	0	0	0
4	595	25	1	3	0	0	0	0	0	0

1 Left 0 Right Phone Dependencies (1-0)

Iteration No.	1	2	3	4-6	7-10	11-20	21-50	51-100	101-1000	>1000
1	124	104	64	121	56	66	52	22	14	1
2	383	147	25	54	8	5	1	1	0	0
3	346	164	23	72	11	6	1	1	0	0
4	348	161	24	73	10	6	1	1	0	0

1 Left 1 Right Phone Dependencies (1-1)

Iteration No.	1	2	3	4-6	7-10	11-20	21-50	51-100	101-1000	>1000
1	124	104	64	121	56	66	52	22	14	1
2	134	116	60	135	52	56	42	19	10	0
3	133	116	60	133	53	56	45	17	11	0
4	134	116	60	136	52	55	43	18	10	0

Figure 2 - Frequencies of number of DP equi-probable path matches for different DP iterations

4.3 Relevance of DP Generated Statistics

The goal of this experiment was to assess the relevance of the insertion, deletion, and substitution probabilities, and study the changes between different iterations. As the TIMIT hand-labeled transcriptions will be used, the effects of co-articulation should be largely accounted for, with any co-articulation effects within the statistics due to effects not adequately represented within the hand-labeled transcriptions.

The experiment was carried out for phone contexts of (0-0) and (1-0). Results are shown for the 1st and 2nd iterations. Figure 3 shows the frequencies with which phone substitutions occurred, as measured after the 1st iteration. Figure 4 shows the frequencies with which the various phone substitutions occurred after the 2nd iteration, permitting a comparison of the differences arising from the use of modified phone confusion probabilities. Due to space constraints only 12 of the 39 phones are displayed in the Figures.

A comparison of the frequency counts shown in Figure 4 to those of Figure 3 clearly shows that the iterative DP match has resulted in a reduced number of low frequency substitutions. This reduction can be attributed to the iterative DP match reducing the number of spurious substitutions arising from equi-probable DP matches. This was also found from an analysis of the complete phone confusion matrix which showed that after the 1st iteration approximately 45% of the total 8,327 phone substitutions were accounted for by 52 significant substitutions, each occurring at least 30 times. For the 2nd iteration a total of 4,844 phone substitutions were observed, with the same 52 significant substitutions now accounting for 63% of all substitutions.

Observed Phone \ Expected Phone	ah	eh	er	ey	ih	d	m	n	r	s	t	z
ah	**549**	59	22	3	120	10	12	10	7	2	1	4
eh	60	**288**	20	29	88	8	6	2	4		4	
er	20	17	**521**	2	32	3		2	108	1		1
ey	1	17	1	**325**	29						1	
ih	166	113	49	76	**1213**	26	10	29	6	5	5	9
d	5	2	1	4	10	**567**	5	9	4	7	41	1
m	4	3	1		9	3	**524**	78	3	4	1	2
n	20	17	4	6	39	62	115	**883**	2	4	7	5
r	10	10	133	6	5	8	5	4	**521**	2	5	1
s	3	2	1	3	5	2	2	3	1	**942**	9	120
t	3	1	3		12	80	1	2		11	**468**	12
z	2	1	1	2	4	2	1	6	1	134	2	**471**

Figure 3 - Phone confusion matrix after the first iteration for a 0-0 context

Observed Phone \ Expected Phone	ah	eh	er	ey	ih	d	m	n	r	s	t	z
ah	**548**	59	14	2	120		6	5	4			
eh	59	**286**	7	24	93	1			2		1	
er	18	13	**521**	1	31	1		1	115			
ey		15		**325**	29							
ih	169	113	26	74	**1205**	5	2	13	3		1	
d				2		**568**	3	8	2		39	
m	1				1	3	**524**	79				
n	9	6		2	29	57	120	**881**		1	2	1
r	6	6	136	2	3	1	4		**522**		2	
s										**945**	7	124
t			1		2	78	1	1		8	**469**	10
z	2				1	1		3		138	2	**473**

Figure 4 - Phone confusion matrix after the 2nd iteration for a 0-0 context

These findings suggest that the most significant substitutions were largely unaffected by the iterative process, whereas the infrequent substitutions (i.e. noise effects arising from undesirable, equi-probable DP matches) were largely removed; the primary goal of the iterative DP match.

5. SUMMARY

The results presented here indicate that the iterative method of refining insertion, deletion and substitution probabilities reduces the number of equi-probable DP matches, to the extent that for the vast majority of utterances only one possible mapping was obtained. Furthermore, it was found that the iterative DP match, when compared to the standard DP match, resulted in the formation of a set of phone confusion probabilities which were more precisely defined and subject to less noise.

Based upon these results, we conclude that the iterated DP match should lead to the formation of improved phonological rules and lexical pronunciation alternatives (and also any other lexical knowledge which is generated through the application of a DP match). Such an improved lexicon will more accurately relate words to their expression in terms of sub-word units, and as a consequence, result in improved speech recognition performance.

6. ACKNOWLEDGEMENTS

This research is supported by the EPSRC under grant GR/K82505. Acknowledgement is also due to British Telecom Laboratories for providing financial support.

7. REFERENCES

[1] Lamel, L. and Adda, G. "On Designing Pronunciation Lexicons for Large Vocabulary Continuous Speech Recognition", *Proceedings of ICSLP-1996*, pp6-9, 1996

[2] Riley, M. "A Statistical Model for Generating Pronunciation Networks", *Proceedings of ICASSP-1991*, Vol. 2, pp737-740, 1991

[3] Imai, T., Ando A. and Miyasaka, E. "A New Method for Automatic Generation of Speaker-Dependent Phonological Rules", *Proceedings of ICASSP-1995*, pp864-867, 1995

[4] Jost, U. Heine, H. and Evermann, G. "What is wrong with the leixcon – An attempt to model pronunciation probabilistically", *Proceedings of Eurospeech-1997*, pp2475-2478, 1997

[5] Kipp, A., Wesenick, M-B. and Schiel F. "Automatic detection and segmentation of pronunciation variants in German speech corpora", *Proceedings of ICSLP-1996*, pp106-109, 1996

[6] Ravishankar, M. and Eskenazi, M. "Automatic Generation of Context-Dependent Pronunciations", *Proceedings of Eurospeech-1997*, pp2467-2471, 1997

[7] Humphries, J.J. and Woodland P.C. "Using Accent-Specific Pronunciation Modeling for Improved Large Vocabulary Continuous Speech Recognition", *Proceedings of ICSLP-1996*, pp2324-2327, 1996

[8] Byrne W., et. al. "Pronunciation Modeling Using a Hand-Labeled Corpus for Conversational Speech Recognition", *Proceedings of ICASSP-1998*, pp313-316, 1998

[9] Sloboda, T. and Waibel, A. "Dictionary Learning for Spontaneous Speech Recognition", *Proceedings of ICSLP-1996*, pp2328-2331, 1996