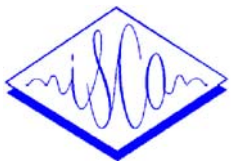

6th European Conference on
Speech Communication and Technology
(EUROSPEECH'99)
Budapest, Hungary, September 5-9, 1999

GENERATING ALTERNATIVE PRONUNCIATIONS FROM A DICTIONARY

Filipp Korkmazskiy and Chin-Hui Lee

Bell Labs, Lucent Technologies
600 Mountain Avenue
Murray Hill, NJ 07974, USA
yelena@research.bell-labs.com, chl@research.bell-labs.com

ABSTRACT

We propose a language independent method for alternative word pronunciation generation using a language specific dictionary. A set of optimal alternative word pronunciations can be generated from a word spelling by using statistically significant associations between strings of letters and strings of phonemes extracted from a dictionary. The proposed method does not require any prior knowledge about the language nor does it need a collection of the speech training data. The alternative pronunciations were used in the recognition experiments. Even though the experiments showed comparable to a baseline system recognition performance they indicated that produced alternative pronuncations could be used for a pronunciation network initialization. This pronunciation network can be further adjusted by a small amount of speech adaptation data. The important advantage of the proposed method is its ability to automatically learn about a language phonological structure. This knowledge can be used while designing complex multilingual systems when information about the languages is limited and speech data for a specific language are unavailable or restricted.

10.21437/Eurospeech.1999-126

1. INTRODUCTION

The existing approaches to pronunciation generation use either language specific pronunciation rules (e.g. [1], [2]) or rules obtained from speech data (e.g. [3], [4], [5], [6]). A selection of language specific rules is usually a subjective one and may not always reflect statistically significant events in the pronunciation domain. The pronunciation rules derived from speech data are dependent on conditions under which the data are collected and in general they are database dependent. Both approaches lack universality in terms of their ability to use for an arbitrary language. Introduced in this paper is an alternative approach to automatic pronunciation rules derivation considering pronunciation modeling as a translation from the letter(source) symbol space into the phoneme(target) symbol space.

Assume we have *a source symbol space and a target symbol space*. The problem is to make a translation of an unknown word from the source symbol space into the target symbol space. The only information available is a pronunciation dictionary that contains a set of examples of the known words presented both in the source symbol space as word spellings and in the target symbol space as the corresponding word pronunciations. Finding *statistically significant correspondencies between the strings of letters and the strings of phonemes* in the word spellings, and the word pronunciations can be considered as a task

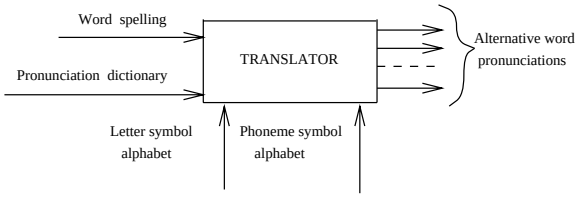

Figure 1: A Pronunciation Generation Problem

of *detection hidden regularities(rules)*. These rules can be used to make a translation from an arbitrary word presented by its spelling to the set of the alternative word pronunciations. Fig. 1 illustrates this problem.

2. ASSOCIATION RELATION

A main notion of our study is *an association relation* that has to be revealed between symbol strings representing some events in the source space and the target space. For example, in a pronunciation modeling task, we wish to find a pronunciation that corresponds to the sequence of letters 'bite'. The only available source of information is a dictionary containing spelling and pronunciation for the different words. For this particular case a solution in the form of a phoneme string 'b ay t', can be found by a dictionary look up. Hereafter, we use the TIMITBET symbols, a superset of ARPABET symbols, for specifying phones [7]. Our task is to find an *automatic method* capable of detecting all such correspondences(associations). Moreover, because the letter symbol strings that could provide nonambiguous associations with the proper phoneme symbol strings are not known we should construct such symbol strings both in the lexical and pronunciation domain in the process of association detection.

Let's provide a mathematical definition for the association relation. A term 'association' relates to a statistical dependency of the random events and can be defined in following way. Let's select a letter symbol string S_l and a phoneme symbol string S_p. An occurrence of the letter string S_l in a word spelling is considered as a random event A and an occurrence of the phoneme string S_p in a word pronunciation is considered as a random event B. Assume a dictionary includes $N(S_p)$ such words that their pronunciations contain the phoneme strings S_p. A probability π of occurrence of the string S_p in a word pronunciation(event B) can be evaluated as follows:

$$\pi = \frac{N(S_p)}{N}. \qquad (1)$$

Here N is a total number of the words in the dictionary. Let's select n such words from the dictionary that their spelling contain the letter string S_l(event A). Assume some k of these n words have the phoneme string S_p included in their pronunciations. In the case of statistical independence of the events A and B a random value k subjects to a binomial distribution (e.g. [8]):

$$P(k,\pi,n) = \binom{n}{k} \cdot \pi^k \cdot (1-\pi)^{(n-k)}. \quad (2)$$

In terms of statistical hypotheses testing *a null hypothesis* means statistical independence for the events A and B. If a value of $P(k,\pi,n)$ gets less than a significance value α we accept *an alternative hypothesis* that claims a statistical dependence for A and B. A mathematical definition for an association relation $A \Rightarrow B$ between random events A and B can be formulated as follows:

$$P(k,\pi,n) < \alpha, \quad and \quad k > \pi \cdot n. \quad (3)$$

Here the probability $P(k,\pi,n)$ is evaluated according to Eq. (2) for binomial distribution and the term $\pi \cdot n$ is an expected value for k. A relation $A \Rightarrow B$ means that an event A influences another event B in such a way that the number k of the instances of the event B dominates over its expected value $\pi \cdot n$ when observing n random events A.

It seems reasonable to define *a strength of an association* in order to be able to make a selection of the best associating events. From the definition for an association relation (3) it is clear that with the probability $P(k,\pi,n)$ decreasing a strength of an association should increase. So, a strength $S(A \Rightarrow B)$ of an association $A \Rightarrow B$ can be defined as follows:

$$S(A \Rightarrow B) = -\log(P(k,\pi,n)). \quad (4)$$

3. ASSOCIATION DETECTION ALGORITHM

Association detection algorithm generates word pronunciation from word spelling by using special knowledge about language phonological structure extracted from a dictionary. It starts from detecting statistically significant correspondencies between string of letters and strings of phonemes in the word spelling and word pronunciation provided in the dictionary. By using a strength of association of the detected correspondencies it aligns spelling and pronunciation for all words in the dictionary and then uses results of alignment for iterative association strength reevaluation. After the pruning the most significant associations are used to translate word spelling into word pronunciation.

3.1. Notations

First of all, let's introduce some notations. *A letter symbol alphabet* $\mathbf{\Lambda_s}$ is a set of the letter symbols and *a phoneme symbol alphabet* $\mathbf{\Lambda_t}$ is a set of the phoneme symbols. A set of the word spellings $\mathbf{V_s}$ includes the ordered strings of the letter symbols from the set Λ_s. A set $\mathbf{V_t}$ of the word pronunciations includes the ordered strings of the phoneme symbols from the set Λ_t. One-to-one correspondence exists between elements of the sets V_s and V_t. A pair (V_s, V_t) represents *a pronunciation dictionary. A letter string alphabet* $\mathbf{\Omega_s}$ is a set of the letter symbol strings which are some parts of the word spellings. *A phoneme string alphabet* $\mathbf{\Omega_t}$ is a set of the phoneme symbol strings which are some parts of the words pronunciations.

We assume that only the sets Λ_s, Λ_t, V_s and V_t are known. Also some regular correspondences(*associations*) exist between strings from the letter string alphabet Ω_s and the strings from the phoneme string alphabet Ω_t. Decoded associations can be used to make a translation of an unknown word from its representation in the letter symbol space to the representation in the phoneme symbol space. We also assume that each string from the letter string alphabet Ω_s may have one or a few associated strings from the phoneme string alphabet Ω_t and vice versa. Let's describe the decoding algorithm.

3.2. Association relation detection

. We have to detect associations between all letter strings consisting of the d letter symbols ($1 \le d \le D$) and all phoneme strings consisting of the f phoneme symbols($1 \le f \le F$). D is a maximum allowed letter string length and F is a maximum allowed phoneme string length. We start from detecting associations between letter strings consisting of a single symbol from the letter alphabet Λ_s and phoneme strings consisting of a single symbol from the phoneme alphabet $\Lambda_t (d = 1, f = 1)$. Then we increment the length of the letter strings($d = d + 1$) and continue the process of detection association between d letter symbol strings and 1 phoneme symbol strings until $d = D$. Then we increment the length of the phoneme strings($f = f + 1$) and continue the process of detection association between f phoneme symbol strings and d letter symbol strings($1 \le d \le D$) until $f = F$. To reduce computational complexity of the association detection algorithm we examine such pairs of d symbol letter strings and f phoneme symbol strings that are composed only of such substrings that have detected association relation.

3.3. Examples

Let's provide some examples illustrating the proposed approach. Suppose we want to determine if there exists a statistically significant association relation between a string of letters 'tal' in a word spelling and a string of phonemes 'tl' in the word pronunciation. For example, the letter string 'tal' is a part of the spelling for words 'coastal', 'glottal', 'mortal', 'postal' and the phoneme string 'tl' is a part of the pronunciations for these words. Let's mark an appearance of the string 'tal' within a word spelling as an event A and an appearance of the string 'tl' within a word pronunciation as an event B. If one wants to check if there exists an association relation between these events(i.e. $A \Rightarrow B$) a criterion (3) should be applied. Assume there are n such words in the dictionary that their spelling include the letter string 'tal'. Some of these words may have the phoneme string 'tl' included in their pronunciation. Let's denote the number of such words as k, ($k \le n$). In the case of an association $A \Rightarrow B$ a value for k becomes high making a conditional probability $\pi(B|A) = \frac{k}{n}$ differ substantially from the unconditional probability $\pi(B)$

According to the previously formulated task conditions we don't have any prior knowledge about the lexical strings and phone strings that have an association relation. So, to reveal the letter string 'tal' and the phoneme string 'tl' as associated events the described above association detection algorithm should be used. According to the algorithm we should detect associations between

single letters and single phonemes($d = 1$): t $\Rightarrow$ t, t $\Rightarrow$ l, a $\Rightarrow$ t, a $\Rightarrow$ l, l $\Rightarrow$ t, l $\Rightarrow$ l(hereafter the symbols on the left hand denote letters and the symbols on the right hand denote phonemes). Then we can detect associations between 2-tuple strings of letters and 1-tuple strings of phonemes($d = 2$): ta $\Rightarrow$ t, ta $\Rightarrow$ l, al $\Rightarrow$ t, al $\Rightarrow$ l. Then we continue for $d = 3$ detecting associations between 3-tuple strings of letters and 1-tuple strings of phonemes: tal $\Rightarrow$ t, tal $\Rightarrow$ l. Finally we detect an association between 3-tuple string of letters and 2-tuple string of phonemes: tal $\Rightarrow$ tl.

3.4. Alignment and pruning

The information about the detected association relations can be used to get an alignment between word spelling and pronunciation. The alignment procedure can be implemented by dynamic programming(DP) method. The strength of association in Eq. 4 can serve as a distance measure for the DP procedure. Once we know segmentation boundaries of the strings in the aligned word spelling and pronunciation we can reevaluate the strength of association more accurately. In turn, the updated association string strength can be used to iterate the alignment procedure with further more accurate reevaluation of the strength of association.

In most practical applications a number of the strings that may have association relation with a specified string is usually bounded. So, a pruning procedure can be applied to exclude the less significant associated strings both in the letter and the phoneme symbol spaces. In the pruning procedure proposed in this study a maximum number of the strings having association relation with a specified string and a minimum value of the association strength are used to eliminate the less significant associated strings. Pruning is implemented for all the letter and phoneme strings.

3.5. Translation

After deriving all association relations between the letter strings and the phoneme strings we can use this information to make *a translation from the letter symbol space to the phoneme symbol space* for a word represented by its spelling. The translation is based on the synthesis of an optimal word pronunciation consisting of the strings from the phoneme symbol space with a maximum value of *an accumulated translation score*. The accumulated translation score $D(W_l \Rightarrow W_p)$ is a sum of the local translation scores $d(S_l(r) \Rightarrow S_p(r))$ between all aligned letter strings $S_l(r)$ and phoneme strings $S_p(r)$ in the word spelling W_l and word pronunciation W_p:

$$D(W_l \Rightarrow W_p) = \sum_r d(S_l(r) \Rightarrow S_p(r)) \qquad (5)$$

The strength of association in Eq. (4) between the letter strings that compose word spelling and the corresponding phoneme strings that compose word pronunciation can serve as a local translation score $d(S_l(r) \Rightarrow S_p(r))$. Dynamic programming technique can be used to make a corresponding selection of the best pronunciation.

4. PRONUNCIATION MODELING EXPERIMENTS

The information about association relations between the letter strings and the phoneme strings can be used to get multiple alternative pronunciations for a word based on its spelling. In preliminary experiments the alternative pronunciations for the words were produced by applying a set of phonological rules to the word baseform pronunciations generated by the Bell Labs text-to-speech(TTS) system [9]. The phonological rules were obtained by detecting associations between lexical and phonological strings. To select between competing alternative pronunciations the strength of association of the applied phonological rules was used as a measure of quality of alternative pronunciations.

In the current study a word dictionary containing about 235000 English words was used to produced a set of phonological rules. The association detection algorithm produced about 2900 phonological rules which represented the possible transformations from the letter strings to the phoneme strings for the letter and phoneme strings consisting of 1 up to 4 symbols. In our experiments, an isolated-word telephone channel speech Phone-Book database [10] was used. Over 1,300 native speakers of American English reflecting different pronunciation styles and dialects were recorded in the database. A set consisting of the 500 words(4200 test samples) from the Phone-Book database was used for recognition.

In a baseline experiment, only baseform pronunciations for each word are used for recognition. 41 context independent HMMs(3 states, 8 mixtures per a state) representing 40 phonemes and a silence model were trained using a different training set database. The word error rate obtained in this experiment was 10.3%. After applying phonological rules we produced multiple alternative pronunciations for the 500 word vocabulary. Table 4 presents some examples of the produced multiple pronunciations(the 1st pronunciation is a baseform pronunciation):

Spelling	Pronunciation
compromise	k ao m p r ax m ay z
	k ao m p r ow m ay z
	k ao m p r aa m ay z
	k ax m p r ax m ay z
mysteriously	m ih s t iy r ie ax s l iy
	m ih s t er ax s l iy
	m ih s t eh r iy ax s l iy
readjustment	r iy ax jh ao s t m ax n t
	r iy ax ih ao s t m eh n t
carbohydrates	k aa r b ow h ay d r ey t s
	k aa r b ow h ay d r ey t eh s
	k aa r b ow h ay d r ae t s

Table 1: Alternative pronunciations generated by means of the dictionary derived rules

Each of the alternative pronunciations in the given example was generated by applying a single rule that modified a proper sequence of the phoneme in the baseform pronunciation. For example, a rule

pro(p r ax, p r aw, p r aa,)

was applied to get the first 2 alternative pronunciations for the word 'compromise'. This rule means that for the letter string 'pro' we can get 3 alternative phoneme strings 'p r ax', 'p r aw' and ' p r aa'. The phoneme string 'p r ax' is presented in the baseform pronunciation 'k ao m p r ax m ay z'. The 2 alternative pronunciations 'k ao m p r ow m ay z' and 'k ao m p r aa m ay z' were generated

by substituting the phoneme string 'p r ax' with alternative phoneme strings 'p r aw' and ' p r aa'. The third alternative pronunciation for the word 'compromise' was generated by applying a rule:

omp(ax m p, ao m p)

for the letter string 'omp'. On the average 1.2 pronunciations per a word were generated for the 500 word vocabulary. The word error rate obtained in this experiment was 10.3%. Even though the experiments showed comparable to a baseline system recognition performance they indicated that produced alternative pronuncations could be used to initialize a pronunciation network when a knowledge about language phonological structure and(or) speech data are unavailable or restricted. This pronunciation network could be improved by using a small amount of adaptation data([3]). Without such an initialization we would need more speech data for a pronunciation network generation.

5. CONCLUSIONS

This paper presents a new technique for alternative pronunciation generation that is based on detection of hidden associations between letter and phoneme strings. This knowledge can be used to make an optimal translation for an arbitrary sequence of the letter symbols to produce the most plausible multiple translations for this sequence. Preliminary experiments have shown an ability of this method to produce multiple pronunciations in the pronunciation modeling task. Recognition experiments with multiple pronunciations generated for a 500 word vocabulary have shown comparable recognition performance between single and multiple word pronunciations. It is believed that this methodology can produce better recognition results if a pronunciation dictionary does not use the pronunciations produced by a TTS system. The knowledge imbedded in TTS systems is dedicated to produce a single word pronunciation and therefore, can not be considered as a good source for multiple pronunciation rule generation. Also pronunciation rules derived from the dictionary-based pronunciations of proper nouns may reveal a greater usefulness in recognition because the variations in pronunciation of proper nouns come mainly from the different ways of unknown letter strings interpretation and production.

A novelty and a strength of the proposed technique is based on an automatic way of composing statistically and phonetically significant groups of symbols both in the lexical and pronunciation domains. An important byproduct of the proposed technique is its ability of producing a high accuracy alignment between the strings of letters and the corresponding strings of phonemes. This alignment may create a set of rules which can be helpful when a complete knowledge about a language phonological structure is unavailable. A lack of knowledge about a phonological structure of a language is a common problem while designing complex multilingual systems. A designer of such a system can take advantage of the automatic rule derivation technique by applying it to supplement his or her knowledge about the language phonological structure. A numerical measure that evaluates a strength of association for the derived rules can help to differentiate between rules in terms of their statistical significance. This algorithm can also be used for pronunciation network initialization when speech data are unavailable or their amount is restricted.

ACKNOWLEDGEMENTS

The authors would like to thank Dr. A. Surendran and Dr. S. Chi-Lin for the discussions and a profitable exchange of ideas during the process of this research.

6. REFERENCES

[1] E. Giachin, A. Rosenberg and C.-H. Lee, "Word juncture modeling using phonological rules for HMM-based continuous speech recognition," *Computer Speech and Language*, **5,** 1991.

[2] M. Wester, J. Kessens, C. Cucchiarini, Helmer Strik, "Modeling Pronunciation Variation: Some Preliminary Results," *Proc. Department of Language and Speech, University of Nijmegen, The Netherlands*, pp. 127-137, **20**, 1996.

[3] F. Korkmazskiy and B.-H. Juang "Statistical Modeling of Pronunciation and Production Variations for Speech Recognition," *Proc. International Conference on Spoken Language Processing, Sydney,* pp. 149-152, December, 1998.

[4] Riley, M., Ljolje, A., Hindle, D., and Pereira, F., "Automatic generation of detailed pronunciation lexicons" *In Automatic Speech and Speaker Recognition: Advanced Topics*, C.-H. Lee, F.K. Soong, and K.K. Paliwal, Eds. Kluwer Academic, Boston, March 1996, ch. 12.

[5] T. Holter and T. Swendsen, "Maximum Likelihood Modeling of Pronunciation Variation," *Proc. Modeling Pronunciation Variation for Automatic Speech Recognition, Rolduc*, pp. 63-66, May, 1998.

[6] N. Gremelie , & J.-P. Martens, "In search of pronunciation rules," *Proc. Modeling Pronunciation Variation for Automatic Speech Recognition, Rolduc,* pp. 23-27, May, 1998.

[7] W. Fisher, V. Zue, J. Bernstein and D. Pallet, "An acoustic-phonetic database," *J. Acoust. Soc. Am.*, **81**, Suppl. 1, 1987.

[8] L.D. Fisher , & G. V. Belle, "Hypothesis Testing for Binomial Distribution," *Biostatistics,* pp. 182-183, 1993.

[9] R. Sproat, editor. *Multilingual Text-To-Speech Synthesis*, Kluwer Academic Publisher, 1998.

[10] J. F. Pitrelli, C. Fong, S.H. Wong, J.R. Spitz, and H.C. Leung, "Phone-Book: A Phonetically-Rich Isolated-Word Telephone-Speech Database," *Proc. International Conference on Acoustics, Speech and Signal Processing, Detroit,* pp. 192-195, May, 1995.