DATA-DRIVEN PRONUNCIATION MODELLING FOR NON-NATIVE SPEAKERS USING ASSOCIATION STRENGTH BETWEEN PHONES

Ingunn Amdal[*]*, Filipp Korkmazskiy and Arun C. Surendran*

Multimedia Communications Research Laboratory
Bell Labs, Lucent Technologies
Murray Hill, NJ 07974, USA
amdal@tele.ntnu.no, {yelena,acs}@research.bell-labs.com

ABSTRACT

In this paper we present an approach to modelling pronunciation variation, particularly for non-native speakers, by modifying the lexicon. In this way we can model several speakers simultaneously, i.e. use the same lexicon and the same acoustic models for all speakers. We use a data-driven approach, i.e. methods based solely on the reference lexicon, the recognizer's acoustic models, and the acoustic data.

We propose a new alignment procedure using an estimated relation measure between the phones in the reference transcription and in the alternative transcription of the new speaker data. This measure discovers statistically significant correspondence between the phones in the two transcriptions. We present this measure as *association strength*. Rules are extracted from the alignment and used to derive pronunciation variants. Following rule pruning based on estimated probability of rules, the most beneficial rules are used to make a common lexicon.

Experiments using the new alignment algorithm on the Wall Street Journal non-native speaker database gave pronunciation rules that performed favourably in comparison to other alignment methods.

1. INTRODUCTION

Automatic speech recognition of non-native speakers with different mother tongues is a difficult task due to the large variation between the speakers. Speaker variation may be captured in the acoustic models or in the lexicon. In this paper we focus on using the lexicon to capture the variation, thus using the same lexicon and the same acoustic models for all speakers. Since modelling of a group of very different speakers using the acoustic models may result in diffuse models, changing the lexicon by pronunciation modelling might give better performance.

Pronunciation variation can be captured using linguistic knowledge, i.e. specific knowledge about how people with different accents pronounce words. This knowledge is not always sufficient for pronunciation modelling. As an example, a transcription of spontaneous American English speech (Switchboard) revealed 80 variants of the word "the" [6]. Non-native speech varies even more, and the phonological rules governing the variation will probably be different for different mother tongues. In such cases a data-driven approach may be more suitable. Such a method can be used independently of the specific database and language, and it can be reused for other tasks without major modifications.

Furthermore, all parts of the recognizer except the lexicon can be optimized with respect to an objective criterion. A data-driven approach will enable us to use the same criteria for the lexicon as for the other parts of the recognizer, making a unified optimization of the whole system possible. We therefore believe a data-driven approach to pronunciation modelling should be preferred.

Generating pronunciation variants using rules that are automatically derived from data is frequently used [2, 4, 8, 12]. We have used a common framework that can be described in four steps:

1. automatically generate an alternative transcription
2. align the reference and alternative transcriptions
3. derive initial rules from the alignment
4. prune the initial rules

This paper focuses on the second step; how to do the alignment in a data-driven and consistent way. As the final goal is to derive a joint lexicon for all speakers, we also present some rule pruning experiments.

The usual approach when aligning the two transcriptions is to use either phonetically based or uniform costs for the phone-to-phone mappings in the dynamic programming algorithm. The phonetic approach uses costs according to knowledge about how pronunciation variation can map one phone to another. As mentioned earlier, such a mapping can be difficult to achieve for a diverse population and a large vocabulary. We propose a new measure to estimate the relations between phones in reference and alternative transcriptions. We call this measure *association strength*, as it uses statistical dependencies to find the association relations between the phones. These associations can be found automatically from the transcriptions, i.e. totally data-driven.

From the alignment we derive rules and combine them, using rule pruning, to make a common lexicon for all speakers. We have tried computing the association strength for

[*]Ingunn Amdal is a research fellow at the Norwegian University of Science and Technology. This work was done while visiting Bell Labs.

each speaker individually and for all speakers combined. This gives two kinds of alignments and rules; individual and common. We have both tried to combine the individual rules to a common lexicon and derive common rules directly from the alignment.

The paper is organized as follows: The association strength is introduced in Section 2, and Section 3 describes the rest of our approach to pronunciation modelling. Section 4 describes the experiments, and the results are presented in Section 5. Conclusions based on these results are presented in Section 6.

2. ASSOCIATION STRENGTH

The association method was first applied in [9] to generate grapheme-to-phoneme rules. We now apply it to find probabilities of phone-to-phone mappings by comparing two sets of phone transcriptions and finding systematic relations. The relations expressed as probabilities of phone-to-phone mappings are used in the alignment of the reference and alternative transcriptions.

There will be differences between the reference transcription and the alternative transcription. The differences that are due to transcription and segmentation errors will not be systematic, but the differences due to pronunciation variation will be. We want an algorithm that can capture all the systematic differences and filter out the chance ones. If, for example, a speaker often substitutes /s/[1] with /z/, we will often see an occurrence of /z/ in the alternative transcription when /s/ is present in the reference transcription. We want the algorithm to assign a high association value to the relation between these two phones. To perform this association, we propose a new method called *association strength* based on statistical co-occurrences of phones.

This can be explained in a statistical framework as hypothesis testing of the mean of a binomial distribution [3]. If the occurrences of a specific phone (event A) in the reference transcription and a specific phone (event B) in the alternative transcription are independent, the events are subjects to the binomial distribution. We can estimate the probability of event B by dividing the number of alternative transcriptions that contains B by the total number of transcriptions. We call this estimate p. If n is the number of occurrences of event A and k is the number of simultaneous occurrences of event A and event B, the probability of the number of occurrences of B given A can be computed using a binomial distribution:

$$P(K = k) = b(k; n, p) = \binom{n}{k} \cdot p^k \cdot (1 - p)^{(n-k)}. \quad (1)$$

When the observed value of k is much higher than the expected value $n \cdot p$, the independence assumption is wrong. The probability of the number of occurrences of B given A using the binomial distribution formula will then be small:

$$k \gg n \cdot p \quad \text{and} \quad P(K = k) = b(k; n, p) < \epsilon \quad (2)$$

[1]We have used the ARPABET phonetic alphabet for the transcriptions.

We therefore use the negative logarithm of this probability as the association strength between event A and B:

$$S(A \Rightarrow B) = -\log[b(k; n, p)]. \quad (3)$$

In order to avoid a high association strength when there is a negative correlation between phones, we need the restriction $k > n \cdot p$.

Which phones are substituted because of pronunciation variation may depend on both the language and the individual speakers. When capturing the phone-to-phone relations from the data we can easily find different sets for different speakers or groups of speakers.

Sorting the phone-to-phone mappings by association strength, we observe that many of the strongest mappings are intuitive. The top three mappings for one of the speakers (4nd) are:
s → z,
z → s, and
p → b.
In this case it seems that the place of articulation is a more important similarity measure than the manner of articulation. For another speaker a high association strength is achieved for the mapping:
l → ow
Although this may seem like an error, a closer look at the alignments revealed that l after ow often was deleted, making this a useful mapping. An example alignment of the reference transcription /ao l s ow/ and the alternative transcription /ow z ow/ for the word "also" is:
(ao → ow)(l → ow)(s → z)(ow → ow).
In this alignment we observe deletion of l and two substitutions; (ao → ow) and (s → z). Without any similarity measure it would be equally probable to get the alignment
(ao → ow)(l → z)(s → ow)(ow → ow),
giving a deletion of s and the substitutions (ao → ow) and (l → z).

We do not want to count relations between phones that are far apart, therefore we have used word segment transcriptions to estimate the association strengths. Using the resulting alignment to restrict the possible phone mappings, we get an iterative procedure to compute the association strengths.

The association strength rely on an estimate of the mean in a binomial distribution. The significance of the measure will be low when we have few samples n to estimate this mean from.

3. PRONUNCIATION RULE DERIVATION

In this section, we will describe our overall system and how the association strength fit into this system.

Alternative transcriptions

The first step in rule generation is finding an alternative transcription that can reveal the true pronunciations of the speakers. The alternative transcriptions can contain two

types of errors. Either they can be too similar to the reference transcription and hide the differences really existing in the data, or they can contain transcription errors.

We have chosen to use a phone loop grammar [8] to make the alternative transcriptions. This will give many transcription errors, but will not be restricted by the reference transcription, except via the acoustic models. As the association strength will help us filter out non-systematic errors, this approach will reveal the systematic differences that we believe are due to pronunciation variation.

In order to maintain timing information we have performed the phone loop recognition on isolated words. This ensures that we compare transcriptions belonging to the same acoustic data. The drawback is that our experiments then are restricted to word internal rules. Some of the variation will appear at word boundaries, and including cross-word rules should be beneficial [5]. Recent studies of non-native speech imply that non-native speakers have less co-articulation between words [14]. The word segmentation will probably contain errors since it is obtained using the reference lexicon and models. However, we have manually inspected some samples and the segmentation for these seems to be satisfactory. We assume that the transcription errors due to segmentation errors will not be systematic and, as for phone loop errors, will be discarded in the rule selection.

The phone loop transcription was performed with 5-best recognition to get more transcriptions. As we have a limited amount of data this is favourable. The transcription parts with highest likelihood occur more often, and the one with lowest likelihood will be changed. An example of this is the 5-best phone loop transcriptions for the word "numerous":

```
/n ow m aa r z/
/n ow m aa r ih z/
/n ow m eh r z/
/n ow m axr z/
/n ow m aa r eh z/.
```

We can be more confident about the first part of this word than the last (the reference transcription is `/n uw m axr ax s/`).

Alignment and rule derivation

Rules representing the pronunciation variation can be extracted from the alignment of the two transcriptions. An alternative to rules is modelling the pronunciations directly from the alternative transcriptions. [4] and [12] achieve this by using initial rules to restrict the alternative transcriptions. A maximum likelihood approach is described in [7]. Except for preliminary experiments we have chosen the rule based approach. The three main reasons for using rule based methods are: 1) Rules depend on smaller segments than words and will occur more often, giving more reliable estimates. This is essential as we have little data and large variation in this task. 2) The vocabulary of the data used for rule derivation can be different from that of the test data. 3) A possible extension to cross-word rules will be easier.

Ideally the rules should capture the difference between the reference pronunciation of a word and the actual pronunciation used by the speaker(s). The context a phone appears in affects the transformation (substitution, deletion, or insertion) of this phone and must be included in the rules. Since in our experiments we have small amounts of data, we have chosen to use only one preceding and one succeeding phone as context for the phone-to-phone rules. Besides, the immediate phone neighbours are shown to be most important [12].

We align the reference and alternative transcriptions for the segmented words using dynamic programming. The cost of a phone-to-phone mapping is set inversely proportional to the probabilities given by the association strength. In this way we use the data to dictate all the parts of the pronunciation modelling. From the alignment we derive context-dependent phone-to-phone mappings, i.e. rules. As we have more phones in the reference than the alternative transcription, the cost in the dynamic programming is set higher for an insertion than for a deletion.

A *rule source segment* consists of the affected phone `A` as well as the two neighbouring phones `x1` and `x2`. If the rule is that `A` can be mapped to `B`, we write this as `x1-A+x2 → B`. We call `B` the *target* of this rule. `B` can be a single phone (substitution of `A` with `B`), several phones (insertion/substitution), or `DELETED` (deletion of `A`). The notation can be explained with an example. The reference lexicon pronunciation of the word "states" is `/s t ey t s/`, while the alternative `/s d ey t s/` is observed in the alternative transcription. The rule for this variation can be written as:
`s-t+ey → d`.

Rule pruning

Most rule based pronunciation techniques need some kind of rule pruning. Adding many variants to the lexicon will increase the confusability and may decrease performance without any assessment of the number of errors corrected compared to the new ones introduced. Two usual approaches are to either retranscribe the adaptation data by forced alignment [4, 12], or to use thresholds based on probabilities for the pronunciations [2, 8]. In this paper we have used several estimates of rule probability.

From the alignment we count all the mappings from each rule source segment to each of the target phone(s) or deletions. From the alignment we also count the occurrences of all rule source segments. An estimate of the rule probability (RPR1) is the ratio between these counts:

$$\hat{P}(\texttt{x1-A+x2 → B}) = \frac{\text{count}(\texttt{x1-A+x2 → B})}{\text{count}(\texttt{x1-A+x2})} \quad (4)$$

As the phone loop transcription contains errors, it can be advantageous to apply some kind of confidence measure to the alignment [8]. One possibility is to limit the rule extraction to the segments with identical source and target context. In this case the estimated rule probability (RPR2) is:

$$\hat{P}(\texttt{x1-A+x2 → B}) = \frac{\text{count}(\texttt{x1-A+x2 → x1-B+x2})}{\text{count}(\texttt{x1-A+x2})} \quad (5)$$

The advantage is that we are more confident that we do not count alignment "errors" as rules, but on the other hand we have fewer occurrences of each rule. An "error" in `x1` or `x2` will wrongly inhibit the rule being counted.

4. EXPERIMENTS

The task: Wall Street Journal non-native speakers

We have applied our method to the Wall Street Journal (WSJ) adaptation and test set for non-native speakers of American English (Spoke3, November 93). This part of the test consists of 10 speakers reading 40 sentences for adaptation and 40–43 sentences for testing. The 10 speakers read the same adaptation sentences, giving a total of 349 different words in the adaptation set, 142 of which were found in the test vocabulary. For the rest of the words we used the lexicon generated for the 64k WSJ vocabulary.

The vocabulary of the test sentences is the 5k WSJ vocabulary. The reference 5k lexicon was generated with the Bell Labs Text-to-Speech system [13]. For the test set a closed trigram language model for the 5k vocabulary was used. A reference recognizer with 12 Mel frequency cepstral coefficients plus log-energy term and their first and second derivatives was trained on 84 native speakers (WSJ0 SI-84). Phonetic decision tree state tying was used to build triphone HMMs with an average of 11 Gaussian pdfs per state [11]. This recognizer gave the baseline result of 29.2% WER. For tests on all 10 speakers the 95% confidence interval for this WER is [28.2 – 30.3]%.

The same acoustic models were used for both segmenting the adaptation sentences by forced alignment, phone loop recognition, and testing. This is done deliberately, because the pronunciation modelling will tailor the lexicon to the present acoustic models. If for any reason the acoustic models are retrained, we may have to regenerate the rules.

Rule derivation

Only rules that were applicable, i.e. where the rule source segment appeared in one or more of the words in the test lexicon, were maintained. We merged the reference lexicon and the new pronunciations, as this is shown to have a positive effect when using small amounts of data to derive rules [4, 8].

We used a threshold of 6 for the number of occurrences of a rule source segment for all rule selection schemes, i.e. count(`x1-A+x2`) $\geqslant 6$. We also used a minimum rule probability threshold of 20% for all rule selection schemes as we wanted to assure that every rule appeared at least twice. The thresholds used in the rule pruning were set identical for all speakers, it might be beneficial to select these individually. We used no pronunciation probabilities in the lexica.

5. RESULTS

Assessment of phone loop transcriptions

We have used insertion penalties in the phone loop to control the number of phone insertions and deletions. Still, the phone error rate is about 50% compared to the reference transcription. (For native speakers we would expect a lower phone error rate.)

As an assessment of the phone loop transcriptions we have derived pronunciations using a maximum likelihood approach similar to the one described in [7]. Instead of allowing all possible candidate pronunciations (i.e. all possible phone sequences of any length), we have restricted the search space to candidate pronunciations obtained by the 5-best phone loop transcription of the word examples. The same method using 10-best phone loop has been shown to give improvements for pronunciation modelling of Norwegian natural numbers [1].

In an initial experiment we used this technique for one of the speakers (4nd). We found speaker dependent pronunciations for the 23 words that occurred more than 3 times for this speaker. 7 of these pronunciations were the same as in the reference lexicon. The WER using these pronunciations for the speaker 4nd was 31.9% compared to the baseline of 34.7%. Although this result is achieved without using the rule based pronunciation modelling the rest of our experiments is based on, it shows that phone loop transcription can be a viable technique for finding pronunciation variants.

Individual rules derived from alignment using association strength

To compare the association strength with other alignment methods, we performed experiments using different cost schemes for the phone-to-phone mappings. We performed alignment using only uniform costs by treating only identity mappings (i.e. the mappings of a phone to itself) different. All non-identity mappings (i.e. the mapping of a phone not to itself) were assigned the same cost. We also used a simple phonological grouping described in e.g. [2], using 4 phone groups; vowels, sonorants, plosives and fricatives.

For the association based alignments we discarded phone-to-phone mappings with low association strength before assigning the costs. The number of non-identity mappings retained was approximately equal to twice the number of phones. Iteration by recomputing the association strength based on the previous alignment converged after about 4 iterations. After the iterations, the number of phone-to-phone mappings increased to about three times the number of phones.

For each speaker we performed alignment using identity, phonological, and association based costs and made individual rules from these alignments. Sorting by estimated rule probability RPR1 $> 50\%$ according to equation (4) gave an overall error rate of 28.7% for the association based lexica. After four iterations of association strength computation, the resulting lexica gave an overall error rate of 28.6% WER. In Figure 1 the results for the different alignment schemes are shown for RPR1 $> 50\%$. As the different speakers have very different WER, the results in Figure 1 are shown as improvement relative to baseline WER. The improvements (deterioration for some speakers) varies, but the association based alignment performs best on average. From Figure 1 we also notice that the pronunciation rules that perform best

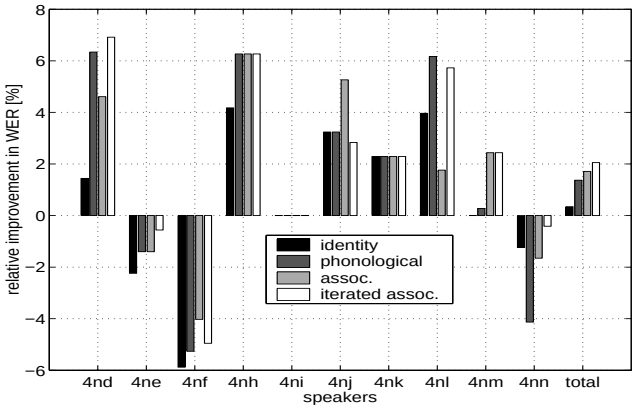

Figure 1: Relative improvement in WER for the different alignment schemes, RPR1 $> 50\%$.

for each speaker are based on one of the association based alignments.

Using equation (5) and a threshold RPR2 $> 20\%$ gave an overall error rate of 28.5%. Using the RPR2 threshold the identity mapping alignment based rules were similar to the association strength alignment.

Common rules for all speakers

We have examined three approaches for deriving rules in order to make a common lexicon for all speakers:

1. derive association strength and alignment for each speaker individually, and derive individual rules and then merge these to a common rule set

2. derive association strength and alignment for each speaker individually, but derive rules for all speakers simultaneously

3. derive association strength, alignment, and rules for all speakers simultaneously

The variation between speakers is large (the speakers have different mother tongues), and it seems reasonable to find rules for each speaker separately and later merge them to make a common lexicon. We wanted to examine the effect of using the association algorithm both on the individual speakers and on all speakers simultaneously. All the experiments for common rule derivations use association strengths after 4 iterations.

To merge the individual rules according to approach 1, we added the counts from each speaker to find a joint rule probability. We used individual rules with RPR1 $> 50\%$ and retained the rules with different estimated joint rule probabilities. For the lexicon with joint RPR1 $> 50\%$, 1559 of the 4986 words in the reference lexicon were affected by the rules, and the average number of pronunciations per word (PPW) was 1.44. In Table 1 results for the different joint rule probability thresholds are shown. We see that the merged lexicon gave larger improvement over baseline than individual lexica. Using the RPR2 scheme did not give the same improvement for the common lexica compared to individual lexica, the best performance was 28.6% WER.

RPR1	*# rules*	*PPW*	*WER*	*rel. improvement*
$> 30\%$	169	1.61	28.4%	2.7%
$> 40\%$	137	1.49	28.2%	3.4%
$> 50\%$	114	1.44	28.6%	2.1%

Table 1: Results for merged rules using different thresholds on the estimated joint rule probability RPR1.

Simultaneously generating rules for all of the speakers from individual association strength derivation and alignment resulted in mostly low probability rules. As the rules are derived using more data, a lower threshold for the rule probability than when deriving rules individually is reasonable. Using a lower threshold gave increased performance over the merged rules, see Table 2. In the RPR1 $> 20\%$ case the recognition was slow because of the huge lexicon.

RPR1	*# rules*	*PPW*	*WER*	*rel. improvement*
$> 20\%$	413	1.84	28.1%	3.8%
$> 30\%$	152	1.30	28.5%	2.4%

Table 2: Results for rules derived for all speakers simultaneously, but with individual association strength derivation, using different thresholds on the estimated joint rule probability RPR1.

Approach 3, generating rules for all of the speakers simultaneously, and also deriving association strength for all speakers simultaneously, gave results similar to approach 2, see Table 3. As the association derivation relies on statistical methods, the individual association derivation may suffer from scarce data. This can be the reason why the association performed for all speakers simultaneously gives a better result. Using the RPR2 scheme gave a WER of 29.0% when generating the rules simultaneously, i.e. hardly any improvement over baseline.

RPR1	*# rules*	*PPW*	*WER*	*rel. improvement*
$> 20\%$	423	1.85	28.0%	4.1%
$> 30\%$	168	1.34	28.4%	2.7%

Table 3: Results for rules derived for all speakers simultaneously using different thresholds on the estimated joint rule probability RPR1.

Relative improvements in WER per speaker for the best lexicon for all the three schemes are shown in Figure 2. We get larger deterioration for one of the speakers using lexica based on common rule derivation than merging individual rules. Besides, the merged lexicon performs similar using fewer rules, 137 compared to 423, which is favourable regarding recognition speed. Testing on native speakers we get a deterioration from 7.8% to 8.3% WER for the merged lexica showed in Figure 2. For both lexica based on simultaneous modelling of all speakers, the native result was 8.5% WER. The merged lexicon is not only smaller but has less confusability as measured on native speakers.

The increased confusability using rules from several speakers in the same lexicon is outweighed by the higher confidence in the rule selection as we use more data. Even if

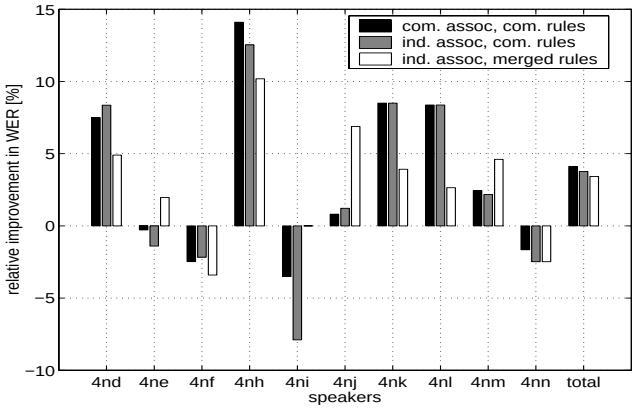

Figure 2: Relative improvement in WER for common lexica using different rule pruning schemes.

the speakers have different mother tongues, similar variation may be present, and we get an improvement by modelling all speakers in the same lexicon. The different schemes presented give different rules and different performance for the different speakers, but the average WER is similar. By using only rule probability measures in the rule pruning we either miss some "good" rules or include some "bad" rules. We have therefore started looking at other rule pruning measures, confident that the association algorithm combined with simple rule probability sorting can give us initial rules which can be further pruned to give larger improvement.

Recent experiments on non-native speakers show larger improvement, from 20.9% to 18.8% WER on another task using re-transcription based on initial context-independent rules [10]. We have not tried re-transcription yet.

6. CONCLUSIONS

In this paper we have presented a completely data-driven approach to modelling a joint lexicon for a group of non-native speakers. We introduced a new metric in the alignment based on statistical co-occurrence called *association strength*. Using the association strength method we can easily find phone relations for individuals or groups of speakers and in this way make the alignments used in the rule generation more consistent.

Lexica modified by rules made from alignments using the association strength measure gave an improvement over baseline that was not present when using uniform phone-to-phone mapping costs in the alignment. This indicates that the proposed method is able to automatically capture the relations in phone variation due to pronunciation variation from comparison of reference and alternative phone loop based alternative transcriptions. The improvement over rules based on phonological alignment was not as high as expected, probably because our statistically based methods suffer from scarce data.

We observed that the lexica generated by merging individual rules performed similar to generating rules for all speakers at the same time, with a WER of 28.2% compared to the baseline result of 29.2%. The merged lexicon scheme gave fewer rules and thus fewer pronunciation per words on average and faster recognition than generating rules for all speakers simultaneously. Testing on native speakers revealed a lower confusability in the merged lexicon.

ACKNOWLEDGEMENTS

The authors would like to thank Olivier Siohan for helpful support on the Bell Labs recognition system.

REFERENCES

[1] Ingunn Amdal, Trym Holter, and Torbjørn Svendsen. Maximum likelihood pronunciation modelling of Norwegian natural numbers for automatic speech recognition. In *Proc. Norwegian Signal Processing Symposium (NORSIG)*, pages 145–150, Asker, Norway, 1999.

[2] Nick Cremelie and Jean-Pierre Martens. In search of better pronunciation models for speech recognition. *Speech Communication*, 29:115–136, 1999.

[3] Lloyd D. Fisher and Gerald van Belle. *Biostatistics*, chapter Hypothesis testing for binomial variables, pages 182–183. Wiley, 1993.

[4] Eric Fosler-Lussier and Gethin Williams. Not just what, but also when: Guided automatic modeling of Broadcast News. In *Proc. DARPA Broadcast News Workshop*, Herndon (VA), USA, 1999.

[5] Egidio P. Giachin, Aaron E. Rosenberg, and Chin-Hui Lee. Word juncture modeling using phonological rules for HMM-based continuous speech recognition. *Computer Speech and Language*, 5:155–168, 1991.

[6] Steven Greenberg. Speaking in shorthand – A syllable-centric perspective for understanding pronunciation variation. *Speech Communication*, 29:159–176, 1999.

[7] Trym Holter and Torbjørn Svendsen. Maximum likelihood modelling of pronunciation variation. *Speech Communication*, 29:177–191, 1999.

[8] Jason J. Humphries and Phillip C. Woodland. Using accent-specific pronunciation modelling for improved large vocabulary continuous speech recognition. In *Proc. EUROSPEECH-97*, pages 2367–2370, Rhodes, Greece, 1997.

[9] Filipp Korkmazskiy and Chin-Hui Lee. Generating alternative pronunciations from a dictionary. In *Proc. EUROSPEECH-99*, pages 491–494, Budapest, Hungary, 1999.

[10] Karen Livescu and James Glass. Lexical modeling of non-native speech for automatic speech recognition. In *Proc. ICASSP-2000*, pages 1683–1686, Istanbul, Turkey, 2000.

[11] Wolfgang Reichl and Wu Chou. Decision tree state tying based on segmental clustering for acoustic modeling. In *Proc. ICASSP-98*, pages 801–804, Seattle (WA), USA, 1998.

[12] Michael Riley, William Byrne, Michael Finke, Sanjeev Khudanpur, Andrej Ljolje, John McDonough, Harriet Nock, Murat Saraclar, Charles Wooters, and George Zavaliagkos. Stochastic pronunciation modelling from hand-labelled phonetic corpora. *Speech Communication*, 29:209–224, 1999.

[13] Richard W. Sproat and Joseph P. Olive. Text-to-speech synthesis. *AT&T Technical Journal*, 74:35–44, 1995.

[14] Laura Mayfield Tomokiyo. Linguistic properties of non-native speech. In *Proc. ICASSP-2000*, pages 1335–1338, Istanbul, Turkey, 2000.