

# Pronunciation Variation Modeling in Automatic Speech Recognition

INGUNN AMDAL AND ERIC FOSLER-LUSSIER



Ingunn Amdal (38) graduated from the Norwegian University of Science and Technology (NTNU) with a *Siv.Ing.* (MSc) in 1989 and a *Dr.Ing.* (PhD) in 2002. In 1990 she started as research scientist at Telenor R&D working on loudspeaking telephones and acoustic echo-cancellation. She joined the newly formed speech technology group at Telenor R&D in 1994 working with automatic speech recognition, dialogue design, and database collection. The topic of her PhD thesis was pronunciation modelling and her current interests include design, representation and evaluation of spoken and multimodal dialogue systems, user tests and spontaneous speech processing.

ingunn.amdal@telenor.com



Eric Fosler-Lussier is currently a Visiting Research Scientist in the Electrical Engineering department of Columbia University, working on novel approaches to automatic speech recognition (ASR) and topic segmentation in recorded multi-party conversations. He received his PhD in 1999 from U. California, Berkeley; his area of dissertation research was dynamic pronunciation modeling for ASR. Subsequently, Fosler-Lussier was a Member of Technical Staff at Bell Labs, Lucent Technologies 2000–2002, conducting research on spoken dialogue system design, natural language call routing, pronunciation modeling, and language modeling for ASR.

fosler@ieee.org

Robust speech recognition is a critical research topic – systems must be able to handle a wide variation in types of speech to make speech technology more user-friendly. One major source of variation in speech is different speaking styles; handling this variation in user input is difficult for current state-of-the-art recognizers. Modeling pronunciation variation within the system can ameliorate the difficulties to some degree. Pronunciation variation can be modeled in different parts of the recognizer; in this presentation we focus on lexical adaptation (other articles in this issue of *Teletronikk* cover other types of robust modeling).

An overview of the methods used in pronunciation variation modeling by lexical adaptation will be given. First, the automatic speech recognition system will be explained briefly with a focus on the pronunciation lexicon. Then, the main distinction between pronunciation modeling methods, knowledge based or data-driven, is explained and illustrated with examples from selected work done in the field. Another distinction often made is modeling of the pronunciation variants directly or indirectly through pronunciation rules that make it possible to generalize knowledge or observations in a training set to unseen data. Finally, a section on confusability reduction is included.

## 1 Introduction

Early automatic speech recognition (ASR) systems only considered restricted speaking styles, i.e. careful articulation of isolated or connected words. The increased modeling capacities of current ASR systems also manage the looser articulation of continuous speech. Making speech technology based applications more widespread has several consequences for the demands on ASR systems:

*New and larger vocabularies are needed when ASR systems are used in new domains.* When the vocabulary is increased it is no longer feasible to select the pronunciation variants by hand. Pronunciations from various sources will often be combined and a decision must be taken whether to include one or more variants per word. The less restricted grammar of a large vocabulary speech recognition system will give more confusability, and more care must be taken in the selection of pronunciation.

*When dialogues between humans and computers are more natural, the ASR must handle more conversational speech.* Conversational speech is harder for ASR systems to recognize correctly, because of increased coarticulation and pronunciation variability, as well as less predictable language usage. Weintraub et al. [1] showed that a spontaneous speaking style is harder to recognize; when the same exact word sequences were recorded in a truly spontaneous, acted spontaneous, and read style, the ASR system performed much worse on spontaneous speech compared with the other two styles.

*Large-scale deployment of systems may increase the amount of dialectal and accented speech that*

*a speech technology application may encounter.*

A more international community increases the portion of non-native speakers in the general public; the substantial differences between native and non-native speech will challenge a natively trained ASR system. Even subtle differences that are easily handled by humans (e.g. Australian versus US English) can still cause problems for ASR.

*The expertise of the user will change the speaking style used.* Novice users will often hyperarticulate, especially when the system misrecognizes them (“Not Austin, I said BOSTON”). Expert users may become more relaxed, and engage with the system using a less formal speaking style. The type of variation seen will depend on the conversational style of prompts in a spoken dialogue system, as well as on the type of task: systems where many people call once or twice will encounter more novice speech, whereas personal dictation systems may encounter speech from experienced users.

Ideally, speech recognizers should handle these diverse speaking styles, (e.g. spontaneous speech, hyperarticulated speech, accents, dialects, and speech from users with different mother tongues). This kind of variation in user input is difficult to model and this is not solved for in current state-of-the-art recognizers.

Pronunciation modeling is by no means a new issue in the ASR community, early efforts are reported in e.g. [2] and [3]. Pronunciation variation modeling is still an important issue in ASR research, and overviews are for example given in [4] and [5]. More recently multilingual ASR has become an interest [6], which introduces new challenges for pronunciation modeling.

## 2 The Automatic Speech Recognition System

The recognition system may be divided into three main elements as shown in the three lines of Figure 1:

- 1 From speech via acoustic features to sub-word units: acoustic models
- 2 From sub-word units to words: lexicon
- 3 From words to sentences: language model

The three modules shown in Figure 1 constitute an automatic speech recognizer. For a recognizer operating in a real speech-based application, the user input will replace the “test speech” in the figure.

Mathematically the system can be described as a classifier. We observe a sequence (i.e. a feature vector representing speech)  $\mathbf{O} = (\mathbf{o}_1, \mathbf{o}_2, \dots, \mathbf{o}_T)$  and want to find the word (sequence)  $W$  that maximizes the *a posteriori probability* (MAP) which will give us the *Bayes classifier*:

$$\hat{W} = \arg \max_W P(W|\mathbf{O}) = \arg \max_W \frac{p(\mathbf{O}|W)P(W)}{p(\mathbf{O})} \quad (1)$$

Since the observation  $\mathbf{O}$  is fixed and  $p(\mathbf{O})$  is independent of  $W$  equation (1) reduces to:

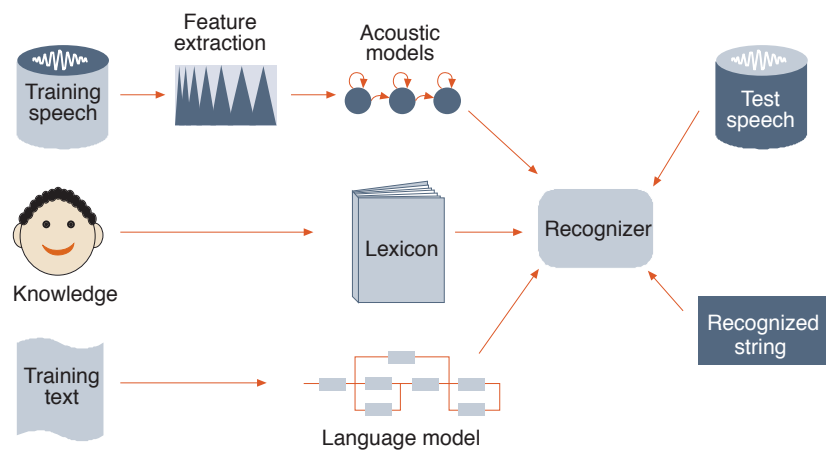
$$\hat{W} = \arg \max_W p(\mathbf{O}|W)P(W) \quad (2)$$

$P(W)$  is generally called the probability of the *language model*, while  $p(\mathbf{O}|W)$  is the probability density of the *acoustic model*.

Recent work [7, 8] has shown that the recognition process can be modeled with a sequence of finite-state transducers (FSTs). An abstract representation of the Viterbi decoding process might be given as:

$$\hat{W} = \text{bestpath}(A \circ T \circ P \circ L) \quad (3)$$

where  $\hat{W}$  is the sequence of words corresponding to the best recognition hypothesis,  $A$  is a finite state automaton (FSA) containing the set of acoustic scores computed from an input utterance,  $T$  is a context-dependent FST, containing a mapping from acoustic states to triphones,  $P$  is the pronunciation model FST, containing a mapping from triphones to words,  $L$  is the language model FSA, which contains  $N$ -gram statistics, and  $\circ$  is the composition operator. All of these finite state machines are typically weighted, with the costs derived from the probabilities of the particular linguistic model.



## 3 How to Model Pronunciation Variation in ASR

Pronunciation variation modeling can be implemented in different parts of the speech recognizer shown in Figure 1. Different realizations of phonemes (allophonic variation) can for example be handled using either more acoustic units, more complex models (but fewer units), or more pronunciation alternatives. The allophonic variation can also be handled by adapting the models to one speaker or a more homogeneous subset of speakers.

The statistically based acoustic models of current ASR systems are capable of handling much of the variation seen in speech, including some pronunciation variation [9]. More complex acoustic models will, for example, handle many allophonic variations in a suitable way. Adaptation of the acoustic models is a successful method to make speaker-dependent recognizers with improved performance compared with speaker-independent recognizers. Task adaptation or adaptation to a group of speakers (e.g. dialect adaptation) is possible, but in general, the more homogeneous the adaptation target, the better the performance. Modeling large variations within the same model by broadening the distributions or adding more components in the Gaussian mixture will give more diffuse, overlapping models that may lead to increased confusability [10]. The success of this type of speaker adaptation also depends on the match between the actual pronunciations and the transcription used in adaptation.

Some of the pronunciation variation is caused by speaking style (dialects, non-native mother tongue, etc.), and may be better handled by careful design of the pronunciation dictionary, i.e. pronunciation modeling [9]. The most common way of dealing with pronunciation variation is to put several pronunciation alternatives in the ASR lexicon. These pronunciations are also often used to retranscribe the speech corpus

Figure 1 Automatic speech recognition system

| Word        | Pronunciation                 |
|-------------|-------------------------------|
| READ        | r eh1 d                       |
| READ (2)    | r iy1 d                       |
| READABILITY | r iy2 d ah0 b ih1 l ih0 t iy0 |
| READABLE    | r iy1 d ah0 b ah0 l           |
| READER      | r iy1 d er0                   |
| READER'S    | r iy1 d er0 z                 |
| REALIZE     | r iy1 l ay2 z                 |
| REALIZE (2) | r iy1 ah0 l ay2 z             |
| RIGHT       | r ay1 t                       |
| TOO         | t uw1                         |
| TWO         | t uw1                         |

Table 1 Examples of ASR lexicon entries

before a retraining of the acoustic models. Using the lexicon to capture speaker variation makes it possible to model several speakers simultaneously, thus using the same lexicon and the same acoustic models for all speakers.

High quality recognizers always include a language model, which is sometimes incorporated in pronunciation modeling techniques. For large vocabulary speech recognition, a well-designed language model may decrease the negative impact of a mismatch between the speaker and the acoustic models and explicit pronunciation modeling may be less important. If the speaking style we try to model has special language model characteristics, e.g. the hesitations and restarts of spontaneous speech, they may be incorporated directly into the language model [11].

One of the main challenges in pronunciation modeling is to know which variation we are attempting to model. The effects of the acoustic models, the lexicon, and the language model will interact, even the choices at the speech pre-processing stage will influence the variation modeling. Superfluous complexity or, even worse, adding contradicting changes, may result from modeling the same variation in several recognizer components. The two main techniques for capturing variation, acoustic model adaptation and lexicon adaptation, should be combined using the method that gives the best result: acoustic model adaptation for the pronunciation variation that can be described as allophonic, and lexicon adaptation for the more phonological variation like deletions and insertions. In this article, we focus on lexicon adaptation.

## 4 The ASR Lexicon

If units other than words are used for the acoustic models, we need a correspondence between the acoustic model units and the words in the vocabulary. When phones form the basic acoustic model units, this corresponds to, but is not necessarily equal to, a pronunciation dictionary. In the ASR community, the term *lexicon* is often used instead of dictionary for this link. The middle line in Figure 1 shows that the lexicon is often based on knowledge contrasting the optimization based on speech data and objective criteria that are used in the other parts of the recognizer. Up till now most ASR systems use only one canonical or a few pronunciations per word, and these pronunciations have typically been transcribed manually.

Some examples of entries in the CMU lexicon [12] are given in Table 1.<sup>1)</sup> A number after the word indicates a non-canonical pronunciation. The numbers after the vowels indicate stress; not all ASR lexica provide this information. Syllable information is also given in some ASR lexica. First of all we note that the word class is usually not indicated. A consequence of this is that there is no distinction between *homographs* (words with the same spelling but different meanings) with different pronunciations, such as the two tenses of “read” in Table 1. Often an ASR lexicon will have only one entry in such cases. For ASR lexica with multiple entries, several pronunciations can be given for one word, e.g. “realize”. There is usually only one entry for words with different senses, but identical spellings and pronunciations (*homonyms*). For example, “right” can mean both a direction and a notion of correctness (in addition to other senses<sup>2)</sup>). Homonymy can be a problem because the word sense information that could be useful for language modeling and semantic parsing is hidden. *Homophones* are words that have different meanings (and usually different spellings), but the same pronunciations. These words will have separate entries, as shown for “two” and “too”, but the ASR system must rely on the language model to resolve which word is recognized. This is the same as for human speech recognition, except that we usually have more contextual information available, such as the setting and theme for the spoken utterance and the identity of the speaker (this is a topic of pragmatics). In ASR, such knowledge is incorporated by using task and dialogue state dependent language models.

<sup>1)</sup> We use the ARPABET [13] for phonetic transcriptions.

<sup>2)</sup> The Concise Oxford Dictionary lists 6 adjectives, 3 adverbs, 4 nouns, and 2 verbs for the word “right”.

When multiple pronunciations are used for each word, pronunciation probabilities may be used to inhibit confusions due to rare pronunciations. Pronunciation probabilities are often defined as part of the language model instead of the lexicon. From a hand-labeled part of the speech database Switchboard<sup>3)</sup> 36 different pronunciations for “the” were found in the test set, and 38 different pronunciations in the training set. Only half of the variants found in the training set were also observed in the test set. The confusability caused by adding all observed variants can also be illustrated by the 35 different words that had the pronunciation [ax] (schwa) [15].

When adding more pronunciations for each word, we can make the word probability dependent on the pronunciation probability. The ASR classifier equation (2) can then be decomposed to include the pronunciations  $\mathcal{B}$  for the word  $W$  (often referred to as the base forms of  $W$ ):

$$p(\mathbf{O}|W)P(W) = \sum_{B \in \mathcal{B}} p(\mathbf{O}|B, W)P(B|W)P(W) \approx \max_{B \in \mathcal{B}} [p(\mathbf{O}|B, W)P(B|W)P(W)] \quad (4)$$

The last line is the Viterbi approximation of using only the best pronunciation. The pronunciation probability  $P(B|W)$  can be defined as a part of the language model and we get the language model probability  $P(B|W)P(W)$ .

It was shown in [16] that augmenting the recognizer lexicon with pronunciation variants found in a general-purpose lexicon gave small performance gains, and most for read native speech. Error analysis showed that the system using a single canonical pronunciation generated different errors than the one using pronunciation variants, although the word error rate was similar. For non-native speech, no improvement was observed for context-dependent acoustic models compared with context-independent models. This speaking style had the largest gain using speaker dependent acoustic model adaptation, but the performance was still far from the results for native speech. For spontaneous speech less improvement was observed by speaker adaptation than for read speech. Even if speaker adaptation was shown to give large improvements, the resulting performance was worse than for native read speech. To achieve results more comparable to native read speech for these two tasks, a combination of lexical and acoustic adaptation may be beneficial.

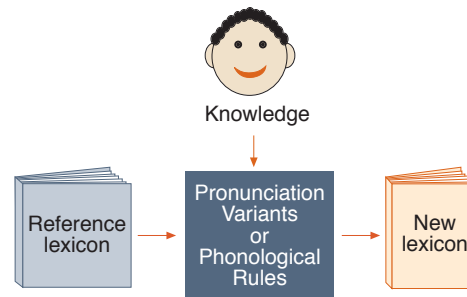


Figure 2 Knowledge-based lexicon adaptation

## 5 Lexicon Adaptation

There are two main directions in finding pronunciation variations, each involving different problems:

- 1 *Knowledge based methods*, where we try to find the best pronunciation rules by applying phonetic and linguistic knowledge as shown in Figure 2. The main problem occurs if the knowledge does not cover the variation we want to model. We may then have too many or too few variations and we may not know how frequent they are.
- 2 *Data-driven methods*, where we use databases of real speech to find the variations present as shown in Figure 3. The problem is that the variations based on a given database may give a result too specific for that database. One of the advantages is that we may compute probabilities for the variants, as opposed to the knowledge-based methods.

For both these methods we can distinguish between *direct* and *indirect* modeling. The pronunciation variants can either be derived directly for each word or indirectly by deriving pronunciation rules and using these rules to generate new pronunciations. Data-driven direct modeling limits us to model only words observed sufficiently many times in the adaptation set, whereas for indirect modeling (both for data-driven and knowledge based) care must be taken

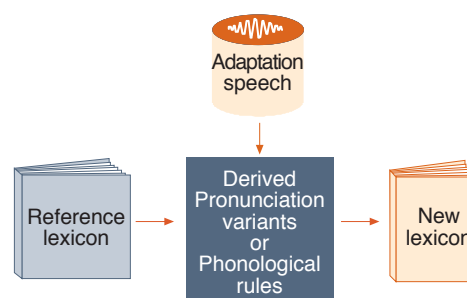


Figure 3 Data-driven lexicon adaptation

<sup>3)</sup> Switchboard is a standard corpus of human-human telephone conversations in US English available from the Linguistic Data Consortium; a portion of it was phonetically transcribed by Greenberg et al. [14].

| Name        | Rule                                      |
|-------------|---|
| Reductions: |   |
| Mid vowels  | -stress [aa ae ah ao eh er ey ow uh] → ax |
| High vowels | -stress [iy ih uw] → ix                   |
| R-vowel     | -stress [er] → axr                        |
| Syllabic n  | [ax ix] n → en                            |
| Syllabic m  | [ax ix] m → em                            |
| Syllabic l  | [ax ix] l → el                            |
| Syllabic r  | [ax ix] r → axr                           |
| Flapping    | [tcl dcl][t d] → dx / V _ [ax ix axr]     |
| Flapping-r  | [tcl dcl][t d] → dx / V r _ [ax ix axr]   |
| H-voicing   | hh → hv / [+voice] _ [+voice]             |

Table 2 Knowledge based phonological rules for US English, after [22]

in the generalization from the observed variation. If a certain variation appears in very different contexts in the adaptation data compared with the test data the generalization may not be valid. Variation observed in function words in the adaptation data may for example not be a variation appropriate for content words in the test data even if the phone context is the same. This can be illustrated by the function word “for” with the canonical pronunciation [f ao r] and the alternative pronunciation [f er]. This transformation is not equally probable for the noun “forest” with the canonical pronunciation [f ao r ah s t].

In the FST formalism given in equation (3), indirect modeling using phonological rules can be represented as an additional phone confusion FST  $C$  between the triphones and the lexicon [17]:

$$\hat{W} = \text{bestpath}(A \circ T \circ C \circ P \circ L) \quad (5)$$

The main reasons to use indirect modeling in a data-driven approach are:

- The vocabulary of the data used for rule derivation can be different from that of the test data. Rules help us generalize the variation seen in the adaptation data to words not present (“unseen words”).
- Rules depend on smaller segments than words and will occur more often, giving more reliable estimates.
- A possible extension to crossword rules will be easier.

Some of the pronunciation variation will be present across word boundaries [18]. Using rules makes the extension to cross-word pronunciation modeling easier, although multi-words make it

possible to model cross-word effects also when dealing directly with pronunciation variants [19]. *Multi-words* are new lexical items formed of several words, e.g. “going to” can be treated as one word to account for the pronunciation variant [g aa n ax].

In general, there has been a migration from knowledge-based methods to data-driven methods. In [20] it was shown that general-purpose lexica do not model spontaneous speech sufficiently well. Only 33 % of the pronunciations found in the hand-labeled part of Switchboard were present in the Pronlex dictionary [21]. The non-canonical pronunciations showed an 11 % increase in word error rate over canonically pronounced words. Another observation was that frequent segments showed more variation, e.g. function words. Data-driven methods will model frequently occurring segments better. This might be an advantage, as frequent words will have a larger influence on the WER. Besides, ASR is based on statistics, and the differences and similarities perceived by humans might not be the most useful for ASR. A combination is thus usual, using some kind of data-driven method to verify the rules and find probabilities for them.

## 5.1 Examples of Knowledge Based Pronunciation Modeling

### 5.1.1 Linguistically Derived Phonological Rules

In [22], Tajchman et al. investigated ten phonological rules for US English. These ten rules are given in Table 2. We observe that the seven reduction rules have no context, whereas the three substitution rules have phone groups as context. Context given as phonetic features or phone groups is common for knowledge based phonological rules. A surface lexicon was built containing the unmodified pronunciations as well as the pronunciations resulting from the application of each relevant phonological rule. Each pronunciation was tagged with the use of rules in order to facilitate counting occurrences of each rule. The main issue in this paper was to estimate the pronunciation probabilities for unseen words using the rule probabilities. The rule probabilities found using automatic transcription (on WSJ) were similar to probabilities found using hand-labeled data on another set (TIMIT). This indicates that the phonological rules have a probability independent of the specific corpus. It was shown that pruning the resulting surface lexicon gave improved recognition results.

Probabilities for linguistically based transformation rules can also be determined by decision trees, as shown by Finken and Waibel [23]. Forced alignment was used to choose among the rules in the training corpus, which was the

Switchboard database (using a speaker adapted recognizer). These data-verified rules were modeled in a decision tree, taking into account the context dependency of phonetic neighbors, word type, speaking rate, average word or phone duration, vowel stress, pitch, and computed probabilities. A rule probability was estimated from the relative frequency of the use of each rule. The resulting transformation rules were interpreted as speaking mode dependent. The lexicon was expanded using pronunciations found by forced alignment, i.e. direct modeling, but with indirect modeling as an intermediate step. The variants found using the derived rules did not increase the performance as much as when selecting variants from the baseline dictionary. The authors interpreted this as due to added confusability. Pronunciation weighting using the speaking mode dependent decision trees increased the performance. The starting point for this experiment was knowledge based, but data were used for verification.

Five known pronunciation rules for Dutch (4 deletion rules and 1 insertion rule) were investigated by Kessens et al. in [24] and Wester et al. in [25]. Improvements were shown by incorporating them in known contexts. Modifying the acoustic models by retranscribing the training data using the variants gave increased improvement. Language model modification was done by incorporating pronunciation probabilities (computed from forced alignment of training data) and gave further improvement. Crossword rules were investigated by including both "border" versions of pronunciations and multi-words. In both cases this was limited to frequently occurring variations. A data-driven approach was compared with this knowledge-based approach by Kessens et al. in [26]. Deletion rules were found by allowing deletions in an alternative transcription, in order to let the acoustic models decide where to delete phones. The knowledge based approach and the data-driven approach gave about the same performance, but the data-driven approach resulted in a smaller lexicon. As the data-driven rule derivation was controlled by frequency counts, the most frequently occurring variations (the most important ones) were favored. There was a 96 % overlap in transcriptions by the two approaches. The data-driven rule context was phone identity, whereas the knowledge rule context contained broader groups of phones. The knowledge-based rules will therefore be applied more often, even if the same transformation (in this case deletion) is described.

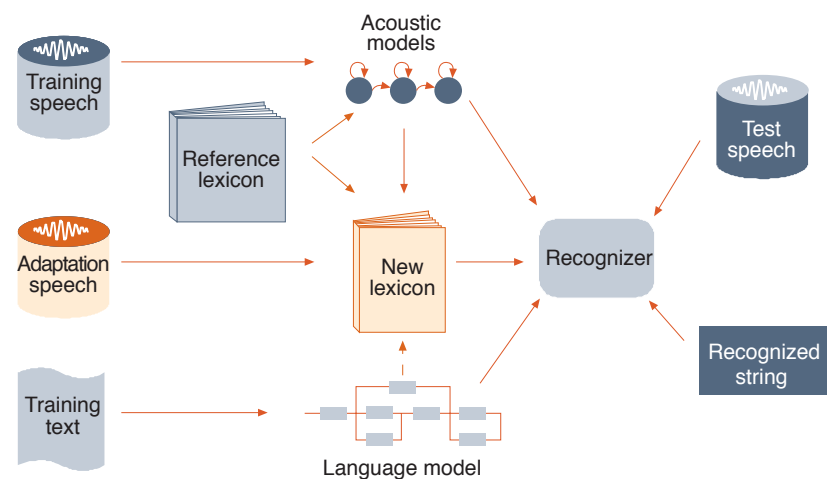
### 5.1.2 Manually Transcribed Data

Using hand-labeled data is also a kind of knowledge-based method, but if the pronunciations found are used in retranscriptions, the categoriz-

ing is less clear-cut. One example of this is experiments performed during the Johns Hopkins summer workshop in 1997 described by Byrne et al. in [27] and Riley et al. in [28]. The results show that pronunciation modeling techniques using automatically labeled data performed better than hand-labeled. The experiments first used indirect pronunciation modeling using the rules to generate variants for unseen words. Decision trees were used to model the phonological rules seen in hand-labeled speech material, i.e. which phones can be neighbors dependent on lexical stress and distance from word boundary. From these decision trees a small network of possible alternate transcription was made for each word. Using all of these alternative pronunciations for recognition decreased the performance. This is the same effect as was shown in [23] and shows the need for some way to choose which pronunciations to accept in the lexicon. The alternative pronunciations from the decision trees were then used in a forced alignment to retranscribe the corpus. Using new decision trees based on this automatic transcription the performance increased. The authors' interpretation was that there was a mismatch between transcription by human perception and machine perception. Another reason may be that the hand-labeled material was much smaller than the automatically transcribed since the process of hand labeling is time consuming and expensive.

Direct modeling with hand-labeled data as bootstrap was also investigated in the experiments of the Johns Hopkins summer workshop in 1997 using an "explicit dictionary expansion", [27] and [28]. Pronunciations found sufficiently often in the hand-labeled or the automatically transcribed corpus were put in the lexicon with weights based on relative frequency. This means fewer variations for the recognizer, and the performance increased, as expected. Multi-words were used to model coarticulation effects be-

Figure 4 Recognition system with adapted lexicon



tween some of the words. The conclusion of the experiments was that crossword context is not necessary except for some words. Using the initial decision tree rules to retranscribe the training set and train new acoustic models increased the performance. These new models were used for a new retranscription, making new decision trees and new explicit lexicon expansions, improved the performance even more. These experiments also showed the need for care when deriving the weights for the alternate pronunciations (i.e. pronunciation probabilities).

## 5.2 Direct Data-driven Pronunciation Modeling

One of the major challenges in pronunciation modeling is to decide which pronunciations to include in the lexicon to get the best ASR performance. Most ASR lexica in use today are based on linguistic knowledge only and are not optimized with respect to ASR performance. Data-driven pronunciation modeling focuses on finding the “best” pronunciations given an objective criterion.

An overview of the recognition system with adapted lexicon based on data-driven methods is shown in Figure 4. The new lexicon is affected by both the adaptation data and the acoustic models, and for most approaches also the knowledge based reference lexicon. For consistency, the language model should also be considered in lexicon adaptation (but often it is not); this is shown by a dashed arrow. The acoustic models may also be adapted by the new data, possibly using the new lexicon. The new lexicon may also influence the language model. For an optimal system the influence between all parts of the system should be considered giving a joint optimization.

Pronunciation variation modeling can be described in two steps:

- 1 Find pronunciation variants
- 2 Assess the variants and modify the lexicon

## 5.3 Examples of Direct Data-driven Pronunciation Modeling

A truly data-driven method was reported by Holter and Svendsen in [29], using some of the methods derived for finding optimal pronunciations presented in [30]. For the experiments a US English database with a 991 word vocabulary was used. No rules or hand-labeled data were used, only a baseline recognizer. This recognizer was used to make an  $N$ -best list of pronunciations without any prior knowledge of the vocabulary other than the number of words and the boundaries of each word, which is usually present in an orthographic transcription. A subset of these pronunciations was chosen through a

maximum likelihood algorithm doing joint optimization for all the utterances of each word. A clustering procedure chose which variants to add in the lexicon to ensure distinct pronunciations, still using the maximum likelihood metric. The appeal of this idea is that it uses objective criteria for the optimization of all parts of the recognizer. The same method, using a 10-best phone loop, has been shown to give improvements for pronunciation modeling of Norwegian natural numbers [31].

Another approach to finding variants using data-driven methods was presented by Fosler-Lussier in [19] and [32]. First, an alternative transcription based on a bigram phone grammar was used to derive rules using decision trees. Then the training set was retranscribed using these rules to get a “smoothed transcription”, and the pronunciation variants that occurred frequently enough were added to the lexicon. An added feature in this approach is the use of “dynamic” lexica, including word frequency, word trigram probability, word length, and speaking rate measures in the pronunciation modeling.

A similar approach was described for the Verbomobil project by Wolff et al. in [33]. A phone recognizer generated alternative hypotheses that were aligned with a transcription based on the canonical lexicon entries. Pronunciations occurring eight times or more were included in the lexicon and gave an increased performance. A measure of confusability called “consolidation” showed that words occurring 15 times or more had reached a “stable” set of pronunciations. This threshold excluded 85 % of the lexicon adaptation material. The algorithm was therefore expanded to incorporate generalization by using frequent sequences instead of words only.

## 5.4 Indirect Data-driven Pronunciation modeling

Generating pronunciation variants using rules that are automatically derived from data is another modeling option. These rules should ideally capture the difference between the reference pronunciation of a word and the actual pronunciation used by the speakers. This approach is similar to the first step in the variant generation of [19] and [33], as well as the approach based on hand-labeled transcriptions in [28].

Decision tree modeling, also called CART (Classification and regression trees) modeling is a popular method for deriving pronunciation rules. It is described in some of the earliest pronunciation modeling approaches, e.g. [3] and [34]. Usually one tree is built for each phone. The number of rules is controlled by limiting the number of mappings in each leaf, and the leaf probabilities can be used as rule probabilities.

Pronunciation variant generation by using data-driven rule derivation can be described in five steps:

- 1 Automatically generate alternative transcriptions
- 2 Align the reference and alternative transcriptions
- 3 Derive rules from the alignment
- 4 Assess and prune the rules
- 5 Generate pronunciation variants from the rules, assess the variants, prune or assign weights, and modify the lexicon

As we can see the first step in direct pronunciation variant modeling is replaced by a 4-step rule derivation and a step 5 to generate variants from the rules. Step 4 is not trivial; rules may interact and one rule may change the context affecting other rules. The rule pruning will control the number of variants indirectly, but we need a step 5 to assess the variants. We may also add a step 6 performing retranscription of the lexicon adaptation material and iterate the process. The five steps are discussed more carefully below:

### Step 1

The first step in rule generation is finding an alternative transcription that can reveal the true pronunciations of the speakers. We can also use knowledge as a starting point: if we have hand-labeled data, pronunciation rules can be derived from comparing this transcription with the reference. The reference transcription will also often be derived semi-automatically. Usually only a word transcription exists, and if the reference lexicon contains several pronunciations, the recognizer is used to choose pronunciation by forced alignment. The third possibility is to use a phone recognizer. As an example, the two transcriptions for the utterance “Paramount Pictures expected eight ...” are shown in Table 3.

### Step 2

The usual approach when aligning the two transcriptions is to use dynamic programming. The difference between the methods lies in how the costs for the phone-to-phone mappings are as-

signed: often either uniform cost or phonetically based costs are used. The latter rely on knowledge about phone similarity and the assumption that the probability of phone-to-phone mappings due to pronunciation variation will follow phone similarities. In [35], an alternative alignment procedure was proposed using an estimated relation measure between the phones in the reference transcription and in the alternative transcription of the new speaker data. This measure utilizes statistically significant correspondence between the phones in the two transcriptions and was called association strength. An example of an alignment of the transcriptions in Table 3 is shown in Table 4.

### Step 3

Rules representing the pronunciation variation can be extracted from the alignment of the two transcriptions. A usual approach in rule based pronunciation modeling is to let the rules express phone-to-phone mappings (allowing deletions and insertions). The rules are usually defined as dependent on a specified context. The width of the context has to be decided as well as which other information to include. The most frequently used context is one phone neighbour to each side of the phone(s) affected by the rule. This is different compared with knowledge-based rules where the context often is given as phonetic features as shown in Table 2. Also for data-driven rules other contexts than phone identity are used and contexts shown to have effect include word frequency [19], lexical stress, and syllabic information [28]. Using more complex contexts demands either more data to estimate the rules properly or a generalization of the contexts. CART trees can be used to generalize the context of a rule automatically. For crossword rules the word boundary information must be included in the context. From the example transcriptions for “Paramount Pictures ...” we can derive several rules. Examples of word internal phone-to-phone-mappings with context given by phone identity are:

[r ah m] maps to [r m] (a deletion)

Reference transcription (canonical pronunciations):

[p eh r ah m aw n t p ih k ch er z ih k s p eh k t ah d ey t]

Alternative transcription (phone loop on spontaneous dictation):

[p eh r m aa m p ih k ch er z ah k s p eh k t ih t ey iy t]

p eh r ah m aw n t p ih k ch er z ih k s p eh k t ah d ey t

p eh r m aa m p ih k ch er z ah k s p eh k t ih t ey iy t

Table 3 Example of reference and alternative phonetic transcriptions for the utterance “Paramount Pictures expected eight ...”

Table 4 Example of alignment of the transcriptions for the utterance “Paramount Pictures expected eight ...”

and

[m aw n] maps to [m aa m] (two substitutions)

In the last case a word-external rule may be more appropriate (\$ marks the word border):

[m aw n t \$ p] maps to [m aa m \$ p]

#### Step 4

Most rule-based pronunciation modeling techniques need some kind of pruning to control which of the alternative pronunciations that should be included in the lexicon. Rarely used pronunciations may introduce more errors than they correct. We can use a threshold based either on the rule probabilities, the pronunciation probabilities, or both, to control the number of new pronunciations to add.

#### Step 5

From the resulting set of selected rules new pronunciations are derived. Using several rules for each pronunciation will result in a huge number of new pronunciations:

$$\begin{aligned} \# \text{ pronunciations} = \\ (\# \text{ rules for phone 1}) \cdot (\# \text{ rules for phone 2}) \dots (6) \end{aligned}$$

The resulting pronunciation probabilities derived from the rule probabilities would be very low for most of the multi-rule derived pronunciations. One way around this is to use estimated pronunciation probabilities as a threshold to limit the number of variants instead of, or in addition to, the rule probability threshold.

Pronunciation probabilities can be estimated directly by counting the different pronunciations chosen when retranscribing the adaptation data by forced alignment (restricted by the initial rules). This was described as “smoothed transcription” in [19] and “explicit” dictionary expansion in [28]. Estimating pronunciation probabilities is not trivial for unseen words. Rule probabilities make it possible to estimate the probability for unseen words and pronunciations, but care must be taken. In [22], the problem of combining the rule probabilities to word probabilities was discussed. Pronunciation probabilities derived from the decision trees were compared with pronunciation probabilities estimated by frequency counts in [28]. These two ways of estimating the pronunciation probability did not give the same result.

When we have little data to derive pronunciations from, it may also be necessary to merge the reference lexicon and the new pronunciations [36]. In [19], merging was found beneficial even if a 100-hour adaptation set was used, because

the pronunciations were modeled directly and the infrequently occurring words got too “noisy” pronunciations.

## 5.5 Examples of Indirect Data-driven Pronunciation Modeling

Humphries and Woodland did experiments on British English accent modeling in [37]. This is an example of indirect modeling, deriving pronunciation rules from data. No hand-labeled transcription was used, so this is an entirely data-driven approach. Alternate transcriptions were found by allowing all vowels to be substituted and then using a forced alignment. Vowel transformations are an important difference between British English accents. The three best transcriptions were used to derive rules using context dependent decision trees including leaf probabilities. A new pronunciation dictionary was then made for the new accent. Typically, an average of 4 pronunciations per word were found to be effective. The test was performed on a 2000 word vocabulary, and adding pronunciations increased the performance.

Humphries and Woodland have also done experiments on British versus US English in [36] and [38]. The recognizer was here used to perform a free (and erroneous) transcription. An acoustic confidence measure was used to filter the transcribed data before rule derivation. This transcription was aligned with the canonical pronunciations in the British English lexicon. A list of possible phone transformations (substitutions, insertions and deletions) was generated and incorporated in a decision tree. Examination revealed interesting correlations with linguistic analysis of the differences between British and US English, e.g. transformation of [t] to [d] in certain contexts. The British-trained recognizer was tested on US English speech using US-adapted pronunciations, and this gave an increased performance compared with using British pronunciations [36]. There was no difference between using British pronunciations and adapted US pronunciations when training new acoustic models on US English speech [38]. The performance was worse but comparable to US models trained on “real” US pronunciations. In the case of sufficient training material, the extra phonological information was of less value. The authors suggested that in this case the pronunciation variation was taken care of by the acoustic models.

An example of the migration from knowledge based to data-driven modeling, is the work of Cremelie et al. Their first work was based on hand-labeled data [39]. Later work in [40] and [41] was based on automatically derived “expert” transcriptions where a constrained speech recognizer was used. In the improved

version [40], more linguistic constraints were put into the automatic transcription, a wider context was used, and negative rules were allowed. Negative rules means that variation is prohibited. In [41], further improvements gave a significant performance gain. Pronunciation rules were derived from alignment of the reference and “expert” transcriptions, and each acoustic segment was only allowed to count for one rule. Rule probabilities were found by frequency counts and the rules were ordered in a rule hierarchy that favored more specific rules. The most important rules were the coarticulation rules between words; most other experiments only consider intra-word variations. These coarticulation rules were incorporated in the language model (since cross-word rules cannot simply be added to the lexicon). For the experiments, they used two databases with different vocabularies for both English and Flemish, i.e. four databases in total. Cross-checking between the two databases for the same language was done to investigate possible corpus-specific rules. The results showed improvement for both languages and about the same level of improvement for rules based on either automatically generated or hand-made transcriptions. The automatic transcription version was best in some tests, the same effect as observed by Byrne et al. [27]. According to the authors, the reason could be that the automatic transcription caught the peculiarities of the recognizer and therefore gave rules better suited for this particular recognizer.

In [42], Amdal et al. used data-driven approaches for all the steps in indirect data-driven pronunciation modeling. For the alignment of the reference and alternative pronunciations the *association strength* introduced in [35] was used to derive phone substitution costs from the data. A metric based on acoustic log likelihood was used in rule pruning. The methods were evaluated on a non-native task. The results showed that the acoustic log likelihood pruning improved the ASR performance compared with the more traditional rule probability pruning. A better performance was observed when modeling the non-native speakers jointly than individually. This was a surprising result, as the speakers had quite different language backgrounds, but may be caused by the small amount of data available. Even if the joint set of non-native speech was more diverse, the larger amount of data was beneficial to get a more reliable rule selection for the data-driven methods investigated.

## 5.6 Confusability Reduction

Confusability reduction is closely linked to the pruning of rules and pronunciation variants. As shown the assessment of pronunciations is usually done by one or a combination of these methods:

- Assess pronunciation rules (the resulting variants will then be assessed indirectly)
- Assess each pronunciation variant directly

Modeling pronunciation variation by adding several pronunciations for each word in the lexicon should be done with care. More variants, and probably more similar variants, will increase the lexical confusability and the error rate. A method of balancing the “old” errors corrected and the “new” errors introduced should be included in the algorithm. Rule and variant assessment only considers the “goodness” of each pronunciation alone and how it performs on the word that the pronunciation belongs to. In general, the pronunciations will interact; thus, a more global approach to assess the total set of pronunciations that takes into account the effect on the other words in the vocabulary should be beneficial. This calls for discriminative techniques incorporating misclassification measures.

Torre et al. explored in [43] measures of word confusability by first estimating a phone confusion matrix. The phone confusions were combined with word confusions that were used in vocabulary selection. If we have several synonyms, this algorithm can be used to choose the best pronunciations among synonyms. With basis in the same algorithms, automatic alternative transcription generation was also examined with promising preliminary tests. Phone confusion matrices were also used by Sloboda and Waibel in [44]. First the confusion matrix was used to smooth a phone bigram that was used to automatically find variants for frequently misrecognized words. For the variants found, homophones as well as variants that only differed in confusing phones, were eliminated.

Discriminative model combination was used for pronunciation modeling experiments by Schramm and Beyerlein in [45]. An expression for the word error count including the pronunciation weights was derived. Minimizing this function with respect to these weights gave an iteration formula for updating the weights. The update function depends on the frequency of occurrence of the pronunciation in the true word as well as for competing words. The technique presented is similar to the one presented by Korkmazskiy and Juang in [46], where discriminative adaptation of pronunciation weights were used.

A confusability measure based on substring matching was presented by Wester and Fosler-Lussier in [47]. All possible word pronunciations that matched substrings in the reference transcription were used to make a lattice of confusable words. The confusability metric was calculated by considering the number of words that

corresponded to each phone. Pronunciations with high confusion counts were then removed.

This method was later extended to incorporate acoustic confusability by Fosler-Lussier et al. in [48]. The FST formulation of the ASR system as shown in equation (3) is utilized by inverting the FSTs. A weighted set of all word sequences  $\mathbf{W}$  confusable with any word sequence  $W$  was derived by composing the given word sequence with inverted transducers until acoustic scores were produced, and then reversing the process:

$$\mathbf{W} = W \circ L \circ P^{-1} \circ T^{-1} \circ T \circ P \circ L \quad (7)$$

A confusion matrix represented by a FST  $C$  was used to model  $T^{-1} \circ T$ . This confusion matrix was trained on the recognition errors of a training set. It was shown that the confusion matrix formalism was able to reasonably predict recognition errors in the test set at least better than chance.

## 6 Concluding Remarks

We have given an overview of pronunciation modeling by lexicon adaptation. There has been a migration from knowledge-based methods to more data-driven methods and combinations of the two.

A use of pronunciation modeling not covered in this article is computer-assisted language learning (CALL) systems. In second-language instruction technologies, pronunciation variation modeling can help capture regular patterns of errors made by students. One challenge in this domain is that the pronunciations of the users (hopefully) will change over time.

For most experiments, the improvements seen are modest. A general observation is that adding variants to the ASR lexicon not only corrects errors, but also introduces errors. One reason for the modest improvements achieved in pronunciation modeling is the lack of a way to control the confusability between pronunciations. To make lexica tailored to a person or group we cannot rely on just adding extra pronunciations, we must also remove confusable ones. One of the main topics in pronunciation variation modeling presently is therefore confusability measures. Speaker dependent lexicon adaptation is also an interesting subject in the current focus of the pronunciation modeling community.

## 7 References

- 1 Weintraub, M et al. Effect on speaking style on LVCSR performance. In: *Proc. ICSLP-96*, Philadelphia (PA), USA, Addendum 16–19, 1996.

- 2 Bahl, L R et al. Large vocabulary natural language continuous speech recognition. In: *Proc. ICASSP-89*, Glasgow, Scotland, 465–467, 1989.
- 3 Riley, M D, Ljolje, A. *Automatic speech and speaker recognition: Advanced topics*, ch. Automatic generation of detailed pronunciation lexicons, 285–301. Kluwer, 1996.
- 4 Strik, H, Cucchiari, C. Modeling pronunciation variation for ASR : overview and comparison of methods. In: *Proc. ESCA Workshop on Modeling Pronunciation Variation for Automatic Speech Recognition*, Rolduc, the Netherlands, 137–144, 1998.
- 5 Strik, H. Pronunciation adaptation at the lexical level. In: *Proc. ISCA ITRW Adaptation methods for speech recognition*, Sophia-Antipolis, France, 123–130, 2001.
- 6 Adda-Decker, M. Towards multilingual interoperability in automatic speech recognition. *Speech Communication*, 35 (1-2), 5–20, 2001.
- 7 Mohri, M et al. Full expansion of context-dependent networks in large vocabulary speech recognition. In: *Proc. ICASSP-98*, Seattle (WA), USA, 665–668, 1998.
- 8 Mou, X, Seneff, S, Zue, V. Context-dependent probabilistic hierarchical sub-lexical modelling using finite state transducers. In: *Proc. EUROSPEECH-2001*, Aalborg, Denmark, 451–455, 2001.
- 9 Jurafsky, D et al. What kind of pronunciation variation is hard for triphones to model? In: *Proc. ICASSP-2001*, Salt Lake City (UT), USA, 577–580, 2001.
- 10 Van Compernelle, D. Recognizing speech of goats, wolves, sheep and ... non-natives. *Speech Communication*, 35 (1-2), 71–79, 2001.
- 11 Stolcke, A, Shriberg, E. Statistical language modeling for speech disfluencies. In: *Proc. ICASSP-96*, Atlanta (GA), USA, 405–408, 1996.
- 12 *CMU Pronunciation Dictionary*. [online], 1998. [cited 2002-03-01]. URL: <http://www.speech.cs.cmu.edu/cgi-bin/cmudict/>
- 13 Lea, W A. *Trends in speech recognition*. NJ, Englewood Cliffs, Prentice Hall, 1980.

- 14 Greenberg, S, Hollenback, J, Ellis, D. Insights into spoken language gleaned from phonetic transcription of the Switchboard corpus. In: *Proc. ICSLP-96*, Philadelphia (PA), USA, Addendum 24–27, 1996.
- 15 McAllaster, D et al. Fabricating conversational speech data with acoustic models : A program to examine model-data mismatch. In: *Proc. ICSLP-98*, Sydney, Australia, 1847–1850, 1998.
- 16 Amdal, I, Svendsen, T. Evaluation of pronunciation variants in the ASR lexicon for different speaking styles. In: *Proc. LREC-2002*, Las Palmas de Gran Canaria, Spain, 1290–1295, 2002.
- 17 Livescu, K, Glass, J. Lexical modeling of non-native speech for automatic speech recognition. In: *Proc. ICASSP-2000*, Istanbul, Turkey, 1683–1686, 2000.
- 18 Giachin, E P, Rosenberg, A E, Lee, C-H. Word juncture modeling using phonological rules for HMM-based continuous speech recognition. *Computer Speech and Language*, 5 (2), 155–168, 1991.
- 19 Fosler-Lussier, E, Williams, G. Not just what, but also when: Guided automatic modeling of Broadcast News. In: *Proc. DARPA Broadcast News Workshop*, Herndon (VA), USA, 171–174, 1999.
- 20 Fosler-Lussier, E, Morgan, N. Effects of speaking rate and word frequency on pronunciations in conversational speech. *Speech Communication*, 29 (2-4), 137–158, 1999.
- 21 *CALLHOME American English Lexicon (PRONLEX)*. [online description], 1995. [cited 2002-03-01]. URL: <http://morph ldc.upenn.edu/Catalog/LDC97L20.html>.
- 22 Tajchman, G, Fosler, E, Jurafsky, D. Building multiple pronunciation models for novel words using exploratory computational phonology. In: *Proc. EUROSPEECH-95*, Madrid, Spain, 2247–2250, 1995.
- 23 Finke, M, Waibel, A. Speaking mode dependent pronunciation modelling in large vocabulary conversational speech recognition. In: *Proc. EUROSPEECH-97*, Rhodes, Greece, 2379–2382, 1997.
- 24 Kessens, J M, Wester, M, Strik, H. Improving the performance of a Dutch CSR by modeling within-word and cross-word pronunciation variation. *Speech Communication*, 29 (2-4), 193–207, 1999.
- 25 Wester, M, Kessens, J M, Strik, H. Pronunciation variation in ASR : Which variation to model? In: *Proc. ICSLP-2000*, Beijing, China, IV:488–491, 2000.
- 26 Kessens, J M, Strik, H, Cucchiari, C. A bottom-up method for obtaining information about pronunciation variation. In: *Proc. ICSLP-2000*, Beijing, China, I:274–277, 2000.
- 27 Byrne, W J et al. Pronunciation modelling using a hand-labelled corpus for conversational speech recognition. In: *Proc. ICASSP-98*, Seattle (WA), USA, 313–316, 1998.
- 28 Riley, M et al. Stochastic pronunciation modelling from hand-labelled phonetic corpora. *Speech Communication*, 29 (2-4), 209–224, 1999.
- 29 Holter, T, Svendsen, T. Maximum likelihood modelling of pronunciation variation. *Speech Communication*, 29 (2-4), 177–191, 1999.
- 30 Holter, T, Svendsen, T. Incorporating linguistic knowledge and automatic baseform generation in acoustic subword unit based speech recognition. In: *Proc. EURO-SPEECH-97*, Rhodes, Greece, 1159–1162, 1997.
- 31 Amdal, I, Holter, T, Svendsen, T. Maximum likelihood pronunciation modelling of Norwegian natural numbers for automatic speech recognition. In: *Proc. Norwegian Signal Processing Symposium (NORSIG)*, Asker, Norway, 145–150, 1999.
- 32 Fosler-Lussier, E. Multi-level decision trees for static and dynamic pronunciation models. In: *Proc. EUROSPEECH-99*, Budapest, Hungary, 463–466, 1999.
- 33 Wolff, M, Eichner, M, Hoffmann, R. Automatic learning and optimization of pronunciation dictionaries. In: *Proc. ISCA ITRW Adaptation methods for speech recognition*, Sophia-Antipolis, France, 159–162, 2001.
- 34 Bahl, L R et al. Decision trees for phonological rules in continuous speech. In: *Proc. ICASSP-91*, Toronto, Canada, 185–188, 1991.
- 35 Amdal, I, Korkmazskiy, F, Surendran, A C. Data-driven pronunciation modelling for non-native speakers using association strength between phones. In: *Proc. ISCA ITRW ASR2000*, Paris, France, 85–90, 2000.

- 36 Humphries, J J, Woodland, P C. Using accent-specific pronunciation modelling for improved large vocabulary continuous speech recognition. In: *Proc. EURO-SPEECH-97*, Rhodes, Greece, 2367–2370, 1997.
- 37 Humphries, J J, Woodland, P C, Pearce, D. Using accent-specific pronunciation modelling for robust speech recognition. In: *Proc. ICSLP-96*, Philadelphia (PA), USA, 2324–2327, 1996.
- 38 Humphries, J J, Woodland, P C. The use of accent-specific pronunciation dictionaries in acoustic model training. In: *Proc. ICASSP-98*, Seattle (WA), USA, 317–320, 1998.
- 39 Cremelie, N, Martens, J-P. Automatic rule-based generation of word pronunciation networks. In: *Proc. EUROSPEECH-97*, Rhodes, Greece, 2459–2462, 1997.
- 40 Cremelie, N, Martens, J-P. In search of better pronunciation models for speech recognition. *Speech Communication*, 29 (2-4), 115–136, 1999.
- 41 Yang, Q, Martens, J-P. Data-driven lexical modeling of pronunciation variations for ASR. In: *Proc. ICSLP-2000*, Beijing, China, I:417–420, 2000.
- 42 Amdal, I, Korkmazskiy, F, Surendran, A C. Joint pronunciation modelling of non-native speakers using data-driven methods. In: *Proc. ICSLP-2000*, Beijing, China, III:622–625, 2000.
- 43 Torre, D et al. Automatic alternative transcription generation and vocabulary selection for flexible word recognizers. In: *Proc. ICASSP-97*, Munich, Germany, 1463–1466, 1997.
- 44 Sloboda, T, Waibel, A. Dictionary learning for spontaneous speech recognition. In: *Proc. ICSLP-96*, Philadelphia (PA), USA, 2328–2331, 1996.
- 45 Schramm, H, Beyerlein, P. Towards discriminative lexicon optimization. In: *Proc. EUROSPEECH-2001*, Aalborg, Denmark, 1457–1460, 2001.
- 46 Korkmazskiy, F, Juang, B-H. Discriminative training of the pronunciation networks. In: *Proc. IEEE Workshop on Automatic Speech Recognition and Understanding*, Santa Barbara (CA), USA, 137–144, 1997.
- 47 Wester, M, Fosler-Lussier, E. A comparison of data-derived and knowledge-based modeling of pronunciation variation. In: *Proc. ICSLP-2000*, Beijing, China, I:270–273, 2000.
- 48 Fosler-Lussier, E, Amdal, I, Kuo, H-K J. On the road to improved lexical confusability metrics. In: *Proc. ISCA ITRW on Pronunciation Modeling and Lexicon Adaptation (PMLA)*, Estes Park (CO), USA, 53–58, 2002.