ELSEVIER

Speech Communication 29 (1999) 225–246

SPEECH COMMUNICATION

www.elsevier.nl/locate/specom

Modeling pronunciation variation for ASR: A survey of the literature

Helmer Strik *, Catia Cucchiarini

A^{2}RT, Department of Language and Speech, University of Nijmegen, P.O. Box 9103, 6500 HD Nijmegen, The Netherlands

Abstract

The focus in automatic speech recognition (ASR) research has gradually shifted from isolated words to conversational speech. Consequently, the amount of pronunciation variation present in the speech under study has gradually increased. Pronunciation variation will deteriorate the performance of an ASR system if it is not well accounted for. This is probably the main reason why research on modeling pronunciation variation for ASR has increased lately. In this contribution, we provide an overview of the publications on this topic, paying particular attention to the papers in this special issue and the papers presented at 'the Rolduc workshop'. [1] First, the most important characteristics that distinguish the various studies on pronunciation variation modeling are discussed. Subsequently, the issues of evaluation and comparison are addressed. Particular attention is paid to some of the most important factors that make it difficult to compare the different methods in an objective way. Finally, some conclusions are drawn as to the importance of objective evaluation and the way in which it could be carried out. © 1999 Elsevier Science B.V. All rights reserved.

Zusammenfassung

Die Forschungsrichtung der automatischen Spracherkennung (ASR) hat sich nach und nach vom Erkennen isolierter Wörter in Richtung Erkennung frei gesprochener Sprache entwickelt. Das hat zur Folge, daß die Aussprachevariation, so wie sie in der freien Rede zutage tritt, bei der Spracherkennung ein intervenierender Faktor geworden ist. Die Leistung eines ASR-Systems wird nämlich erheblich beeinträchtigt, wenn man diesen Faktor nicht berücksichtigt. Dies ist vermutlich der Hauptgrund dafür, warum die systematische Berücksichtigung der Aussprachevariation bei der ASR in letzter Zeit stark zugenommen hat. Dieser Artikel stellt einen Überblick der Literatur zu diesem Thema dar, wobei den Beiträgen in diesem 'special issue' sowie denen des 'Rolduc workshop' besondere Aufmerksamkeit geschenkt wird. Zunächst werden die wichtigsten Unterschiede der zahlreichen Arbeiten zur Modellbildung der Aussprachevariation diskutiert. Dann folgt eine Besprechung der Beurteilung und des Vergleichs verschiedener Methoden, die der Modellbildung zugrunde liegen. Dabei wird den wichtigsten Faktoren, die einen objektiven Vergleich der Methoden erschweren, besondere Aufmerksamkeit geschenkt. Letztendlich schließen sich einige Schlußfolgerungen im Hinblick auf die Relevanz objektiver Beurteilung und deren mögliche Realisierung an. © 1999 Elsevier Science B.V. All rights reserved.

* Corresponding author. Tel.: +31-24-3616104; fax: +31-24-3612907.

E-mail address: strik@let.kun.nl (H. Strik)

[1] Whenever we mention 'the Rolduc workshop' in the text we refer to the ESCA Tutorial and Research Workshop "Modeling pronunciation variation for ASR" that was held in Rolduc from 4 to 6 May 1998. This special issue of *Speech Communication* contains a selection of papers presented at that workshop.

0167-6393/99/$ - see front matter © 1999 Elsevier Science B.V. All rights reserved.
PII: S0167-6393(99)00038-2

Résumé

Le centre d'intérêt dans la recherche de la reconnaissance automatique de la parole (ASR), parti des mots isolés, s'est engagé vers le discours conversationnel. Par conséquence, la quantité de variation de prononciation présente dans le discours dont nous rapportons les résultats a graduellement augmenté. La variation de prononciation détériorera la performance d'un système ASR si l'on n'en rend pas compte. C'est probablement la raison principale pourquoi la recherche dans le domaine de la modélisation de la variation de prononciation pour ASR a augmenté récemment. Dans cette contribution on fournit une vue d'ensemble des publications sur ce sujet, et en particulier on réfère aux articles de cette edition spéciale et aux contributions présentées dans les sessions qui ont eu lieu a 'Rolduc'. D'abord, les caractéristiques les plus importantes qui distinguent les diverses études sur modélisation de variation de prononciation sont discutées. Puis les questions d'évaluation et de comparaison sont adressées. Une attention particulière est prêtée à certains des facteurs les plus importants qui rendent difficile de comparer les différentes méthodes d'une maniere objective. Enfin quelques conclusions sont tirées quant à l'importance de l'évaluation objective et de la façon dans laquelle elle pourrait être effectuée.

Keywords: Pronunciation variation; Automatic speech recognition

1. Introduction

If words were always pronounced in the same way, automatic speech recognition (ASR) would be relatively easy. However, for various reasons words are almost always pronounced differently. The most important sources of pronunciation variation will be discussed here.

A first major distinction in pronunciation variation can be drawn between intraspeaker and interspeaker variations. Intraspeaker variation refers to the fact that the same speaker can pronounce the same word in different ways depending on various factors. The first important factor that may affect the way in which words are pronounced is the fact that "they are strung together into connected speech" (Kaisse, 1985, p. 1) as opposed to when they are pronounced in isolation. In connected speech, all sorts of interactions may take place between words, which will result in the application of various phonological processes such as assimilation, co-articulation, reduction, deletion and insertion. The degree to which these phenomena occur will vary depending on the style of speaking the speaker is engaged in (for an overview of the literature on speaking styles see (Eskenazi, 1993)). Stylistic variations are usually interpreted as variations in the degree of formality of speech (Labov, 1972). As speech becomes less formal, the syllabic structure of words may be reorganized, speech rate may increase, and there may be changes in pitch and loudness (Laver, 1994, pp. 66–69). Besides stylistic variation there is also free variation, in which the speaker is free to choose from among different pronunciations of the same word, without this having any implications for speaking style.

Another important source of variation in speech is the interlocutor, since it is known that speakers are influenced by the person they are talking to (Coupland, 1984; Giles and Powesland, 1975; Giles and Smith, 1979). In certain accounts of language variation, "the interlocutor" is not viewed as a separate factor, but is incorporated in the style dimension, since style is considered as the "speakers' response to their audience" (Bell, 1984, p. 145).

The types of variation mentioned so far can all take place in the speech of one and the same speaker, though the degree to which they occur is likely to vary between speakers. In addition to this, there is variation in pronunciation between speakers since speakers of the same language may speak different dialects or speak with a different accent (Laver, 1994, pp. 55–56). The specific dialect or accent of a speaker will depend on factors such as region of origin, socioeconomic background, level of education, sex, age, group membership and mother tongue (an overview of such factors, and their effect on pronunciation, is provided in (Scherer and Giles, 1979)). In this respect, it is important to note that intraspeaker variation and interspeaker variation (in the sociolinguistic literature often referred to as stylistic and social variation, respectively) are not independent of each other (Labov, 1972; Romaine, 1980). On the contrary, according to Bell (1984, p. 158) "intraspeaker variation derives from and echoes interspeaker variation".

Over and above the variation due to the factors mentioned so far, which can be called linguistic variation, there is also variation caused by anatomical differences between speakers, such as differences in vocal tract length, variation due to environmental factors such as background noise (Lombard effect) and variation caused by paralinguistic factors such as emotional status (joy, anger, sorrow, etc.) (Polzin and Waibel, 1998). A good review of the literature on human vocal emotion is given in (Murray and Arnott, 1993).

Owing to all these sources of variation, each word in a language can be pronounced in many different ways, which constitutes a major problem for ASR. In the beginning of ASR research, the amount of pronunciation variation was limited by using isolated words. In isolated word recognition, the speakers have to pause between words, which of course reduces the degree of interaction between words. Moreover, in this case speakers also have the tendency to articulate more carefully. Although using isolated words makes the task of an ASR system easier, it certainly does not do the same for the speaker, because pausing between words is highly unnatural. Therefore, attempts were made in ASR research to improve technology, so that it could handle less artificial speech. Consequently, the type of speech used in ASR research has gradually progressed from isolated words, via connected words and carefully read speech, to conversational or spontaneous speech. Although many current applications still make use of isolated word recognition (e.g., dictation), in ASR research the emphasis is now on spontaneous or conversational speech. It is clear that in going from isolated words to conversational speech the amount of pronunciation variation increases. Since the presence of variation in pronunciation may cause errors in ASR, modeling pronunciation variation is seen as a possible way of improving the performance of the current systems.

The fact that pronunciation variation should be accounted for in ASR was already noted in the early 1970s. For instance, in many articles in the proceedings of the "IEEE Symposium on Speech Recognition" from April 1974 (Erman, 1974) it is mentioned that multiple pronunciations should be present in the lexicon and that phonological rules can be used to generate them (Barnett, 1974; Cohen and Mercer, 1974; Friedman, 1974; Jelinek et al., 1974; O'Malley and Cole, 1974; Oshika et al., 1974; Rabinowitz, 1974; Rovner et al., 1974; Shockey and Erman, 1974; Tappert, 1974). In the last decade, there has been an increase in the amount of research on this topic, which is evident from the growing number of contributions to conferences (see e.g. Strik, 1998), from the organization of the Rolduc workshop (Strik et al., 1998), and also from the appearance of this special issue of *Speech Communication*.

In spite of the importance that this topic has acquired, it seems that giving a precise definition of pronunciation variation modeling for ASR is not easy. Strictly speaking, one could say that almost all ASR research is about modeling pronunciation variation. As a matter of fact, the ubiquitous "hidden Markov models" (HMMs) are a way of accounting for segmental and temporal variation. In order to better account for coarticulation effects, HMMs have been further refined and made specific for the various contexts, the context-dependent HMMs. Furthermore, the use of multiple Gaussian mixtures has been introduced as a better way of modeling segmental variation. By now, these techniques have become standard, and they are no longer considered as ways of modeling pronunciation variation. In general, when speaking about pronunciation variation modeling for ASR, one thinks of techniques other than these standard ones, as appears from a review of the papers that are presented at conferences under this heading. It follows that it is difficult to say where standard ASR ends and pronunciation variation modeling for ASR begins.

Similarly, it is difficult to define the type of pronunciation variation that has been modeled for the purpose of ASR. First of all, this variation cannot be characterized in terms of the categories that are usually applied in (socio)linguistic research. In general, it can be stated that when the term pronunciation variation is used within the context of ASR, it usually refers to the types of linguistic variation mentioned above under the categories of intraspeaker and interspeaker variation. Although linguistic variation takes place both at the segmental and at the suprasegmental level, in general only segmental variation has been modeled so far in ASR. Furthermore, variation due to environmental characteristics and paralinguistic factors, normally is not explicitly modeled. Interspeaker variation is sometimes (partly) modeled by using speaker adaptation (e.g., anatomical differences are modeled with vocal tract length normalization techniques). Second, the choice of what should be modeled is based on the phenomena observed in the speech material and on the possible effects of these phenomena on recognition performance. For instance, researchers choose to model French liaison, /n/-deletion or voicing assimilation regardless of the factors that may have caused these processes. In many cases, researchers do not even choose which processes to model, they just emerge from analyses of the speech material. This means that it is often difficult to give a precise definition of the type of variation that is modeled in a specific approach, because in most cases it is a combination of different types. In turn, this makes it difficult to compare the various approaches, and to interpret the results of this kind of research from the point of view of human speech processing.

A survey of the literature on pronunciation variation modeling in ASR reveals that most of these papers are directly concerned with testing a specific method for variation modeling to determine whether it leads to an improvement in recognition performance. In addition, there have been studies that were specifically aimed at getting more insight into pronunciation variation so as to develop better approaches to modeling it in ASR (e.g., Adda-Decker and Lamel, 1998, 1999; Fosler-Lussier and Morgan, 1998, 1999; Greenberg, 1998, 1999; Peters and Stubley, 1998; Strik and Cucchiarini, 1998). However, studies of this kind are less frequent.

In this paper we intend to give an overview of the various approaches to modeling pronunciation variation that have been proposed so far, paying particular attention to the papers contained in this special issue and to those presented at the Rolduc workshop (Strik et al., 1998). Where necessary, reference will be made to related research that has been presented elsewhere. Providing such an overview turned out to be difficult because not all authors present their work with the same perspective and the same degree of detail. For instance, the majority of the papers on this topic presented in the reference section of the present article, are papers that have appeared in proceedings of conferences and workshops. Since the size of such proceedings papers is always limited, it is inevitable that details will be missing.

The presentation of the various methods in Section 2 will be organized around some of the major characteristics that distinguish pronunciation variation modeling techniques from each other. After having presented the different techniques (in Section 2), we will address the issues of evaluation and comparison (in Section 3), which are crucial if we want to draw conclusions as to the merits of the various proposals. In particular, we will discuss the most important factors that make it difficult to compare the different methods in an objective way.

2. Characteristics of the methods

As explained above, it is difficult to classify the type of pronunciation variation modeled in ASR in terms of the categories that are usually applied in linguistics and phonetics. That is why we decided to adopt another framework for classification. The framework we will use is based on the decisions that are made when one has to choose for a method for modeling pronunciation variation. These decisions concern the following questions:

1. Which type of pronunciation variation should be modeled?
2. Where should the information on variation come from?
3. Should the information be formalized or not?

4. In which component of the automatic speech recognizer should variation be modeled?

It is obvious that each of these questions cannot be answered in isolation. On the contrary, the answers will be highly interdependent. Depending on the decision taken for each of the above questions, different methods for pronunciation variation modeling can be distinguished, as will appear from the following sections. For each question it is possible to identify a specific dimension along which a choice can be made. In this way a descriptive framework can be obtained to classify the various contributions to modeling pronunciation variation in ASR. Although it is certainly possible that the extremes of some of these dimensions will not occur in practice, this is irrelevant since their main function is to provide us with a framework for description.

2.1. Type of pronunciation variation

The majority of the contributions are concerned with variation at the segmental level. A common way of describing segmental pronunciation variation in the context of ASR is by indicating whether it refers to word-internal or to cross-word processes, because this choice is strongly related to the properties of the speech recognizer being used. As a matter of fact, the choice for word-internal variation, cross-word variation or both, is determined by factors such as the type of ASR, the language, and the level at which modeling will take place.

Modeling within-word variation is an obvious choice if the ASR system makes use of a lexicon with word entries, because in this case variants can simply be added to the lexicon. Given that almost all ASR systems use a lexicon, within-word variation is modeled in the majority of the methods (Adda-Decker and Lamel, 1998, 1999; Aubert and Dugast, 1995; Bacchiani and Ostendorf, 1998, 1999; Beulen et al., 1998; Blackburn and Young, 1995, 1996; Bonaventura et al., 1998; Cohen and Mercer, 1975; Cremelie and Martens, 1995, 1997, 1998, 1999; Ferreiros et al., 1998; Finke and Waibel, 1997; Fosler-Lussier and Morgan, 1998, 1999; Fukada and Sagisaka, 1997; Fukada et al., 1998, 1999; Heine et al., 1998; Holter, 1997; Holter and Svendsen, 1998, 1999; Imai et al., 1995; Kessens and Wester, 1997; Kessens et al., 1999; Lamel and Adda, 1996; Lehtinen and Safra, 1998; Mercer and Cohen, 1987; Mirghafori et al., 1995; Mokbel and Jouvet, 1998; Ravishankar and Eskenazi, 1997; Riley et al., 1998, 1999; Ristad and Yianilos, 1998; Schiel et al., 1998; Sloboda and Waibel, 1996; Svendsen et al., 1995; Torre et al., 1997; Wester et al., 1998a; Williams and Renals, 1998; Zeppenfeld et al., 1997).

Besides within-word variation, cross-word variation also occurs, especially in continuous speech. Therefore, cross-word variation should also be accounted for. A sort of compromise solution between the ease of modeling at the level of the lexicon and the need to model cross-word variation is to use multi-words (Beulen et al., 1998; Finke and Waibel, 1997; Kessens et al., 1999; Nock and Young, 1998; Pousse and Perennou, 1997; Ravishankar and Eskenazi, 1997; Riley et al., 1998; Sloboda and Waibel, 1996; Wester et al., 1998a). In this approach, sequences of words (usually called multi-words) are treated as one entity in the lexicon (see also Section 2.4.1) and the variations that result when the words are strung together are modeled by including different variants of the multi-words. It is important to note that, in general, with this approach only a small portion of cross-word variation is modeled, e.g., that occurring between words that figure in very frequent sequences. Besides the multi-word approach, other methods have been proposed to model cross-word variation such as (Aubert and Dugast, 1995; Blackburn and Young, 1995, 1996; Cohen and Mercer, 1975; Cremelie and Martens, 1995, 1997, 1998, 1999; Mercer and Cohen, 1987; Perennou and Brieussel-Pousse, 1998; Pousse and Perennou, 1997; Safra et al., 1998; Schiel et al., 1998; Wiseman and Downey, 1998).

Given that both within-word and cross-word variation occur in running speech, it will probably be necessary to model both of them. This has already been done in (Beulen et al., 1998; Blackburn and Young, 1995, 1996; Cohen and Mercer, 1975; Cremelie and Martens, 1995, 1997, 1998, 1999; Finke and Waibel, 1997; Kessens et al., 1999; Mercer and Cohen, 1987; Riley et al., 1998, 1999;

Schiel et al., 1998; Sloboda and Waibel, 1996; Wester et al., 1998a).

2.2. Information sources

Another feature that distinguishes the various approaches to modeling pronunciation variation in ASR is the source from which information on pronunciation variation is derived. In this connection, a distinction can be drawn between data-driven versus knowledge-based methods. The major difference between these two types of approaches is that in the former case the assumption is that the information on pronunciation variation has to be obtained in the first place. In knowledge-based approaches, on the other hand, it is assumed that this information is already available in the literature.

The idea behind data-driven methods is that information on pronunciation variation has to be obtained directly from the signals (Bacchiani and Ostendorf, 1998, 1999; Blackburn and Young, 1995, 1996; Cremelie and Martens, 1995, 1997, 1998, 1999; Fosler-Lussier and Morgan, 1998, 1999; Fukada and Sagisaka, 1997; Fukada et al., 1998, 1999; Greenberg, 1998, 1999; Heine et al., 1998; Holmes and Russell, 1996; Holter, 1997; Holter and Svendsen, 1998, 1999; Imai et al., 1995; Mirghafori et al., 1995; Mokbel and Jouvet, 1998; Nock and Young, 1998; Peters and Stubley, 1998; Polzin and Waibel, 1998; Ravishankar and Eskenazi, 1997; Riley et al., 1998, 1999; Ristad and Yianilos, 1998; Sloboda and Waibel, 1996; Svendsen et al., 1995; Torre et al., 1997; Williams and Renals, 1998). To this end, the acoustic signals are analyzed in order to determine all possible ways in which the same word or phoneme is realized. A common stage in this analysis is transcribing the acoustic signals. Subsequently, the transcriptions can be used for different purposes, as will be explained in Sections 2.3 and 2.4. Transcriptions of the acoustic signals can be obtained either manually (Cremelie and Martens, 1995, 1997; Downey and Wiseman, 1997; Fosler-Lussier and Morgan, 1998, 1999; Greenberg, 1998, 1999; Heine et al., 1998; Mirghafori et al., 1995; Riley et al., 1998, 1999; Ristad and Yianilos, 1998; Wiseman and Downey, 1998) or (semi-)automatically (Adda-Decker and Lamel, 1998, 1999; Bacchiani and Ostendorf, 1998, 1999; Beulen et al., 1998; Cremelie and Martens, 1997, 1998, 1999; Fukada and Sagisaka, 1997; Fukada et al., 1998, 1999; Holter, 1997; Kessens and Wester, 1997; Kessens et al., 1999; Lehtinen and Safra, 1998; Mokbel and Jouvet, 1998; Ravishankar and Eskenazi, 1997; Riley et al., 1998, 1999; Schiel et al., 1998; Wester et al., 1998a; Svendsen et al., 1995; Torre et al., 1997; Williams and Renals, 1998). The latter is usually done either with a phone(me) recognizer (Fukada and Sagisaka, 1997; Fukada et al., 1998, 1999; Mokbel and Jouvet, 1998; Ravishankar and Eskenazi, 1997; Torre et al., 1997; Williams and Renals, 1998) or by means of forced recognition (Adda-Decker and Lamel, 1998, 1999; Bacchiani and Ostendorf, 1998, 1999; Beulen et al., 1998; Cremelie and Martens, 1997, 1998, 1999; Kessens and Wester, 1997; Kessens et al., 1999; Lehtinen and Safra, 1998; Riley et al., 1998, 1999; Schiel et al., 1998; Wester et al., 1998a).

In forced recognition (which is also called forced alignment), the ASR system can only choose between the pronunciation variants of a word, and not between all words present in the lexicon, as is the case for a "normal" ASR system. Consequently, forced recognition can be employed to decide which pronunciation variant best matches the signal, and in this way a new transcription can be obtained (see also Section 2.4.2). The performance of forced recognition has been evaluated in (Wester et al., 1998a,b, 1999), by comparing it with the performance of humans that carried out the same tasks. It turned out that for the tasks studied in (Wester et al., 1998a,b, 1999), the performance of the human listeners and forced recognition were similar, and that on average, the degree of agreement between ASR system and listeners is only slightly lower than that between listeners. Therefore, forced recognition seems to be a suitable tool for obtaining information on pronunciation (variation). However, since in (Wester et al., 1999) it is shown that the agreement depends on the properties of the ASR system used, one should be cautious in applying such a tool.

Given that making manual transcriptions is extremely time-consuming, and therefore costly, it is not feasible to obtain manual transcriptions of large corpora. As a consequence, the use of automatically obtained transcriptions is becoming more common. Moreover, there is another reason why transcriptions obtained automatically with the ASR system itself could be beneficial, viz., that these transcriptions are more in line with the phone strings obtained later during recognition with the same ASR system. This is also mentioned by Riley et al. (1998, 1999).

In knowledge-based studies, information on pronunciation variation is primarily derived from sources that are already available (Adda-Decker and Lamel, 1998, 1999; Aubert and Dugast, 1995; Bonaventura et al., 1998; Cohen and Mercer, 1975; Downey and Wiseman, 1997; Ferreiros et al., 1998; Finke and Waibel, 1997; Kessens and Wester, 1997; Kessens et al., 1999; Kipp et al., 1996; Kipp et al., 1997; Lamel and Adda, 1996; Lehtinen and Safra, 1998; Mercer and Cohen, 1987; Mouria-Beji, 1998; Nock and Young, 1998; Perennou and Brieussel-Pousse, 1998; Pousse and Perennou, 1997; Roach and Arnfield, 1998; Safra et al., 1998; Schiel et al., 1998; Wesenick, 1996; Wester et al., 1998a; Wiseman and Downey, 1998; Zeppenfeld et al., 1997). The existing sources can be pronunciation dictionaries (see e.g., Roach and Arnfield, 1998) and linguistic studies on pronunciation variation. However, these sources usually only provide information as to the form of the possible variants, while quantitative information on the frequency of the alternative variants still has to be obtained from the acoustic signals, as is the case for data-driven methods. Furthermore, probably not many suitable pronunciation dictionaries do exist.

The distinction between data-driven and knowledge-based is related to that between bottom–up and top–down, which are also commonly used terms in ASR literature. However, in this paper these terms will not be used interchangeably. More explicitly, in our taxonomy the terms data-driven and knowledge-based are taken to refer to *the starting point* of the research, be it the acoustic signals (data) or the literature (knowledge). On the other hand, the terms bottom–up and top–down refer to *the direction* of the developing process, which can be upward or downward.

Although many studies contained in this issue are not completely data-driven or knowledge-based, it is generally possible to say whether the starting point of the research was mainly data-driven or knowledge-based (see the references above). However, most of them cannot be said to be completely bottom–up or top–down, because in none of these studies is the direction of the developing process solely upward or downward, but the flow of information is in both directions. For example, in many data-driven studies the results of the bottom–up analyses are used to change the lexicon and the altered lexicon is then used during recognition in a top–down manner. Similarly, knowledge-based methods are usually not strictly top–down, e.g. because in many of them the rules applied to generate pronunciation variants may be altered on the basis of information derived from analysis of the acoustic signals.

In general terms, it is not possible to say whether a data-driven study is to be preferred to a knowledge-based one. A possible drawback of knowledge-based studies is that there could be a mismatch between the information found in the literature and the data for which it has to be used. In the introduction it was stated that many current systems are designed for spontaneous speech. However, the knowledge on pronunciation variation that can be found in the literature usually concerns other speaking styles. Therefore, it is possible that the information obtained from the literature does not cover the type of variation in question, whereas information obtained from data could be more effective for this purpose. This form of mismatch between the knowledge and the data may lead to overcoverage, i.e., the addition of variants that do not figure in the corpus, or to undercoverage, i.e., the exclusion of variants that do figure in the corpus. To overcome these problems, one can resort to a combination of top–down and bottom–up approaches, as explained above.

On the other hand, a possible drawback of data-driven studies is that for every new corpus and/or ASR system the whole process of transcribing the speech material and deriving information on pronunciation variation has to be started again. In other words, information obtained on the basis of data-driven studies does not easily generalize to situations other than the one in question. Moreover, the problem of undercoverage may also arise in data-driven approaches, if the corpus is not representative enough.

A good option might be to use a method consisting of two stages. In the first stage, a knowledge-based approach is used, which has the advantage that it can easily be ported to a new task. In the second stage, a data-driven approach is used to model (part of) the remaining pronunciation variation. The data-driven approach should also be used to test existing knowledge and acquire new knowledge. In this way, the amount of pronunciation variation modeled in the first stage will gradually increase, and the importance of the second stage will gradually diminish.

2.3. Information representation

Regardless of whether a data-driven or a knowledge-based approach is used, it is possible to choose between formalizing the information on pronunciation variation or not. In general, formalization means that a more abstract and compact representation is chosen, e.g., rewrite rules or artificial neural networks.

In a data-driven method, the formalizations are derived from the data (Cremelie and Martens, 1995, 1997, 1998, 1999; Deshmukh et al., 1996; Fukada and Sagisaka, 1997; Fukada et al., 1998, 1999; Imai et al., 1995; Ravishankar and Eskenazi, 1997; Torre et al., 1997). In general, this is done in the following manner. The bottom–up transcription of an utterance is aligned with its corresponding top–down transcription obtained by concatenating the transcriptions of the individual words contained in the lexicon. Alignment is done by means of a dynamic programming (DP) algorithm (Cremelie and Martens, 1995, 1997, 1998, 1999; Fosler-Lussier and Morgan, 1998, 1999; Fukada and Sagisaka, 1997; Fukada et al., 1998, 1999; Heine et al., 1998; Ravishankar and Eskenazi, 1997; Riley et al., 1998, 1999; Torre et al., 1997; Williams and Renals, 1998; Wiseman and Downey, 1998). The resulting DP-alignments can then be used to:

- derive rewrite rules (Cremelie and Martens, 1995, 1997, 1998, 1999; Ravishankar and Eskenazi, 1997);
- train decision trees (Fosler-Lussier and Morgan, 1998, 1999; Riley et al., 1998, 1999),
- train an artificial neural network (ANN) (Deshmukh et al., 1996; Fukada and Sagisaka, 1997; Fukada et al., 1998, 1999);
- calculate a phone confusion matrix (Torre et al., 1997).

In these four cases, the information about pronunciation variation present in the DP-alignments is formalized in terms of rewrite rules, decision trees, ANNs and a phone confusion matrix, respectively.

In a knowledge-based approach, formalized information on pronunciation variation can be obtained from linguistic studies in which rules have been formulated. In general, these are optional phonological rules concerning deletions, insertions and substitutions of phones (Adda-Decker and Lamel, 1998, 1999; Aubert and Dugast, 1995; Cohen and Mercer, 1975; Ferreiros et al., 1998; Finke and Waibel, 1997; Kessens et al., 1999; Lamel and Adda, 1996; Lehtinen and Safra, 1998; Mercer and Cohen, 1987; Nock and Young, 1998; Perennou and Brieussel-Pousse, 1998; Pousse and Perennou, 1997; Safra et al., 1998; Schiel et al., 1998; Wester et al., 1998a; Wiseman and Downey, 1998; Zeppenfeld et al., 1997). The formalizations (either obtained from data or from linguistic studies) can then be used to generate surface forms (pronunciation variants) from the base forms.

The obvious alternative to using formalizations is to use information that is not formalized, but enumerated. Again, this can be done either in a data-driven or in a knowledge-based manner. In data-driven studies, the bottom–up transcriptions can be used to list all pronunciation variants of one and the same word. These variants and their transcriptions (or a selection of them) can then be added to the lexicon. Alternatively, in knowledge-based studies it is possible to add all the variants of one and the same word contained in a pronunciation dictionary. Quite clearly, when no formalization is used, it is not necessary to generate the variants because they are already available.

It is not easy to determine a priori whether formalized information will work better than enumerated information. It may at first seem that using formalizations has two important advantages. First, one has complete control over the process of variant generation. At any moment it is possible to select variants automatically in different ways. Second, since the information on pronunciation variation is expressed in more abstract terms, it follows that it is not limited to a specific corpus and that it can easily be applied to other corpora. Both these operations will be less easy with enumerated information. On the other hand, the use of formalizations also has some disadvantages, like overgeneration and undergeneration, owing to incorrect specifications of the rules applied (Cohen, 1989, p. 51). Both types of problems should not arise when using enumerated information.

2.4. Level of modeling

Given that the recognition engines of most ASR systems consist of three components, there are three levels at which variation can be modeled: the lexicon, the acoustic models, and the language model. This is not to say that modeling at one level precludes modeling at one of the other levels; on the contrary, to obtain a good recognition system, it is necessary that concerted modeling happens on the three levels. Therefore, in most studies modeling takes place at more than one level. Nevertheless, in order to categorize the various studies, each category will be discussed separately in the following subsections.

2.4.1. Lexicon

At the level of the lexicon, pronunciation variation is usually modeled by adding pronunciation variants (and their transcriptions) to the lexicon (Adda-Decker and Lamel, 1998, 1999; Aubert and Dugast, 1995; Beulen et al., 1998; Bonaventura et al., 1998; Cohen and Mercer, 1975; Cremelie and Martens, 1995, 1997, 1998, 1999; Downey and Wiseman, 1997; Ferreiros et al., 1998; Finke and Waibel, 1997; Fukada et al., 1998, 1999; Holter, 1997; Holter and Svendsen, 1998, 1999; Imai et al., 1995; Kessens and Wester, 1997; Kessens et al., 1999; Lamel and Adda, 1996; Lehtinen and Safra, 1998; Mercer and Cohen, 1987; Mokbel and Jouvet, 1998; Nock and Young, 1998; Ravishankar and Eskenazi, 1997; Riley et al., 1998, 1999; Roach and Arnfield, 1998; Sloboda and Waibel, 1996; Torre et al., 1997; Wester et al., 1998a; Williams and Renals, 1998; Wiseman and Downey, 1998; Zeppenfeld et al., 1997). The rationale behind this procedure is that with multiple transcriptions of the same word the chance is increased that for an incoming signal the speech recognizer selects a transcription belonging to the correct word. In turn, this should lead to lower error rates.

However, adding pronunciation variants to the lexicon usually also introduces new errors because the acoustic confusability within the lexicon increases, i.e., the added variants can be confused with those of other entries in the lexicon. This can be minimized by making an appropriate selection of the pronunciation variants, by, for instance, adding only the set of variants for which the balance between solving old errors and introducing new ones is positive. Therefore, in many studies tests are carried out to determine which set of pronunciation variants leads to the largest gain in performance (Cremelie and Martens, 1995, 1997, 1998, 1999; Fukada et al., 1998, 1999; Holter, 1997; Holter and Svendsen, 1998, 1999; Imai et al., 1995; Kessens and Wester, 1997; Kessens et al., 1999; Lehtinen and Safra, 1998; Mokbel and Jouvet, 1998; Nock and Young, 1998; Riley et al., 1998, 1999; Sloboda and Waibel, 1996; Torre et al., 1997; Wester et al., 1998a). For this purpose, different criteria can be used, such as:

- frequency of occurrence of the variants (Kessens and Wester, 1997; Kessens et al., 1999; Ravishankar and Eskenazi, 1997; Riley et al., 1998, 1999; Schiel et al., 1998; Wester et al., 1998a; Williams and Renals, 1998),
- a maximum likelihood criterion (Holter, 1997; Holter and Svendsen, 1998, 1999; Imai et al., 1995),
- confidence measures (Sloboda and Waibel, 1996), and
- the degree of confusability between the variants (Sloboda and Waibel, 1996; Torre et al., 1997).

A description of a method to detect confusable pairs of words or transcriptions is also given in

(Roe and Riley, 1994). If rules are used to generate pronunciation variants, then certain rules can be selected (and others discarded), as in (Cremelie and Martens, 1995, 1997, 1998, 1999; Lehtinen and Safra, 1998; Schiel et al., 1998) where rules are selected on the basis of their frequency and application likelihood.

As was mentioned earlier, multi-words can also be added to the lexicon, in an attempt to model cross-word variation at the level of the lexicon. Optionally, the pronunciation variants of multi-words can also be included in the lexicon. By using multi-words Beulen et al. (1998) and Wester et al. (Kessens et al., 1999; Wester et al., 1998a) achieve a substantial improvement, while for Nock and Young (1998) this was not the case.

Before variants can be selected, they have to be obtained, in the first place. Sometimes the pronunciation variants are generated manually (Aubert and Dugast, 1995; Riley et al., 1998, 1999) or selected from enumerated lists (Flach, 1995), but usually they are generated automatically by means of various procedures:

- rules (Adda-Decker and Lamel, 1998, 1999; Aubert and Dugast, 1995; Cohen and Mercer, 1975; Cremelie and Martens, 1995, 1997, 1998, 1999; Flach, 1995; Kessens and Wester, 1997; Kessens et al., 1999; Mercer and Cohen, 1987; Nock and Young, 1998; Ravishankar and Eskenazi, 1997; Schiel et al., 1998; Wester et al., 1998a),
- ANNs (Fukada and Sagisaka, 1997; Fukada et al., 1998, 1999),
- grapheme-to-phoneme converters (Lehtinen and Safra, 1998),
- phone(me) recognizers (Mokbel and Jouvet, 1998; Nock and Young, 1998; Ravishankar and Eskenazi, 1997; Sloboda and Waibel, 1996; Williams and Renals, 1998),
- optimization with maximum likelihood criterion (Holter, 1997; Holter and Svendsen, 1998, 1999) and
- decision trees (Fosler-Lussier and Morgan, 1998, 1999; Riley et al., 1998, 1999).

In (Aubert and Dugast, 1995; Beulen et al., 1998) the transcriptions of the variants of function words and foreign names are generated manually, while the variants for all other words are generated by rule.

Since rule-based methods are probably used most often, it is interesting to note that Nock and Young (1998) conclude that (for their English task) "rule-based learning methods may not be the most appropriate for learning pronunciations when starting from a carefully constructed, multiple pronunciation dictionary". The question here is whether this conclusion is also valid for other applications in other languages, and whether it is possible to decide in which cases the starting point is a carefully constructed, multiple pronunciation dictionary (see also Section 3).

Given that there are various ways to obtain pronunciation variants, one might wonder what the optimal way is. So far not many studies have been reported in which different methods were compared. An exception is (Flach, 1995), in which two types of methods for obtaining variants, rule-based and enumerated, are compared. The baseline system makes use of a canonical lexicon with 194 words. If the variants generated by rule are added to the canonical lexicon, making a total of 291 entries, a substantial improvement is observed. However, if all variants observed in the transcriptions of a corpus are added to the canonical lexicon, making a total of 897 entries, an even larger improvement is found. In this particular example, adding all variants found in the corpus would seem to produce better results than adding a smaller number of variants generated by rule. In this respect, some comment is in order.

First, in this example the number of entries in the lexicon was small. It is not clear whether similar results would be obtained with larger lexica. One could imagine that confusability does not increase linearly, and with many entries and many variants it could lead to less positive results.

Second, the fact that a method in which variants are taken directly from transcriptions of the acoustic signals works better than a rule-based one could also be due to the particular nature of the rules in question. As was pointed out in Section 2.2, rules taken from the literature are not always the optimal ones to model variation in spontaneous speech, while information obtained from data may be much better suited for this purpose.

2.4.2. Acoustic models

Pronunciation variation can also be represented at the level of the acoustic models, for instance by optimizing the acoustic models (Aubert and Dugast, 1995; Bacchiani and Ostendorf, 1998, 1999; Beulen et al., 1998; Bonaventura et al., 1998; Deng and Sun, 1994; Finke and Waibel, 1997; Godfrey et al., 1997; Greenberg, 1998, 1999; Heine et al., 1998; Holter, 1997; Kessens and Wester, 1997; Kessens et al., 1999; Lamel and Adda, 1996; Mirghafori et al., 1995; Nock and Young, 1998; Riley et al., 1998, 1999; Schiel et al., 1998; Sloboda and Waibel, 1996; Wester et al., 1998a). Optimization can be attained in different ways, as will be discussed in the current section.

An obvious way of optimizing the acoustic models is by using forced recognition. In Section 2.2 we already explained how forced recognition can be employed to calculate new transcriptions of the signals. In turn, the new transcriptions can be used to train new acoustic models. These new acoustic models can then be used to do forced recognition again, etc. In other words, this process can be iterated. We will refer to this procedure as iterative transcribing. Forced recognition and iterative transcribing have been used often to obtain improved transcriptions and improved acoustic models (Aubert and Dugast, 1995; Bacchiani and Ostendorf, 1998; Bacchiani and Ostendorf, 1999; Beulen et al., 1998; Finke and Waibel, 1997; Kessens and Wester, 1997; Kessens et al., 1999; Riley et al., 1998, 1999; Schiel et al., 1998; Sloboda and Waibel, 1996; Wester et al., 1998a).

In order to evaluate these procedures, the error rates obtained with the new acoustic models can be compared to those obtained with the old acoustic models. In general, these procedures seem to improve the performance of the ASR systems (Aubert and Dugast, 1995; Bacchiani and Ostendorf, 1998, 1999; Finke and Waibel, 1997; Kessens and Wester, 1997; Kessens et al., 1999; Riley et al., 1998, 1999; Schiel et al., 1998; Sloboda and Waibel, 1996; Wester et al., 1998a). However, Beulen et al. (1998) found that in some cases the performance does not improve, but remains unchanged or even deteriorates. In our own research (Kessens and Wester, 1997; Kessens et al., 1999; Wester et al., 1998a), we found that for iterative transcribing the improvement during the first iteration is almost always (much) larger than that obtained during the following iterations, so usually one iteration is sufficient.

In forced recognition, pronunciation variants are present in the lexicon during training in order to train new acoustic models. Optionally, pronunciation variants can be retained in the lexicon during recognition (testing). In general, using the variants during recognition is more beneficial than using variants during training, while the best results are obtained when multiple variants are included during both training and recognition (Kessens and Wester, 1997; Kessens et al., 1999; Lamel and Adda, 1996; Wester et al., 1998a). Therefore, it seems worthwhile to test the procedure of forced recognition because it is a relatively straightforward procedure that can be applied almost completely automatically and because it usually gives an improvement over and above that of using multiple variants during recognition only.

Optimizing the existing acoustic models is one way in which one could try to improve the performance of an ASR system. Another way is by searching for other (hopefully better) basic units. In most ASR systems, the phone is used as the basic unit and, consequently, the lexicon contains transcriptions in the form of strings of phone symbols. However, in some studies experiments are performed with basic units of recognition other than the phone.

For this purpose, sub-phonemic models have been proposed (Deng and Sun, 1994; Godfrey et al., 1997). In (Deng and Sun, 1994) a set of multi-valued phonological features is used. First, the feature values of the speech units in isolation are defined followed by the (often optional) spreading of features for speech units in context. On the basis of the resulting feature-overlap pattern a pronunciation network is created. The starting point in (Godfrey et al., 1997) is a set of symbols for (allo-) phones and sub-phonemic units. These symbols are used to model pronunciation variation due to context, coarticulation, dialect, speaking style and speaking rate. The resulting descriptions (in which almost half of the segments are optional) are used to create pronunciation networks. In both cases, the ASR system decides during decoding what the optimal path in the pronunciation networks is.

Besides sub-phonemic models it is also possible to use basic units larger than phones, e.g., (demi-) syllables (Greenberg, 1998, 1999; Heine et al., 1998) or even whole words. It is clear that using word models is only feasible for tasks with a limited vocabulary (e.g., digit recognition). For most tasks the number of words, and thus the number of word models to be trained, is simply too large. Therefore, in some cases, word models are only trained for the words occurring most frequently, while for the less frequent words sub-word models are used. Since the number of syllables is usually much smaller than the number of words (Greenberg, 1998, 1999; Heine et al., 1998), the syllable would seem to be suited as the basic unit of recognition. Greenberg (1998, 1999) mentions several other reasons why, given the existing pronunciation variation, the syllable is a suitable candidate.

If syllable models are used, the within-syllable variation can be modeled by the stochastic model for the syllable, just as the within-phone variation is modeled by the acoustic model of the phone (see e.g., Heine et al., 1998). For instance, in phone-based systems deletions, insertions and substitutions of phones have to be modeled explicitly (e.g., by including multiple pronunciations in the lexicon), while in a syllable-based system these processes would result in different realizations of the syllable.

In most ASR systems, the basic units are defined a priori. Furthermore, while the acoustic models for these basic units are calculated with an optimization procedure, the pronunciations in the lexicon are usually handcrafted. However, it is also possible to allow an optimization procedure to decide what the optimal pronunciations in the lexicon and the optimal basic units (i.e., both their size and the corresponding acoustic models) are (Bacchiani and Ostendorf, 1998, 1999; Holter, 1997). In both (Bacchiani and Ostendorf, 1998, 1999) and (Holter, 1997) the optimization is done with a maximum likelihood criterion.

In (Greenberg, 1998, 1999) no tests are described. For the syllable models in (Heine et al., 1998) the resulting levels of performance are lower than those of standard ASR systems. Furthermore, in (Bacchiani and Ostendorf, 1998, 1999; Deng and Sun, 1994; Godfrey et al., 1997; Holter, 1997) the observed levels of performance are comparable to those of phone-based ASR systems (usually for limited tasks). Although these results are interesting, it remains to be seen whether these methods are more suitable for modeling pronunciation variation than standard phone-based ASR systems, especially for tasks in which a large amount of pronunciation variation is present (e.g., for conversational speech).

2.4.3. Language models

Another component in which pronunciation variation can be taken into account is the language model (LM) (Cremelie and Martens, 1995, 1997, 1998, 1999; Deshmukh et al., 1996; Finke and Waibel, 1997; Fukada et al., 1998, 1999; Kessens et al., 1999; Lehtinen and Safra, 1998; Perennou and Brieussel-Pousse, 1998; Pousse and Perennou, 1997; Schiel et al., 1998; Wester et al., 1998a; Zeppenfeld et al., 1997). This can be done in several ways, as will be discussed below.

Let X be the speech signal that has to be recognized. The goal during decoding then is to find the string of words W that maximizes $P(X|W) * P(W)$. Usually N-grams are used to calculate $P(W)$. If there is one entry for every word in the lexicon, the N-grams can be calculated in the standard way. As we have seen above, the most common way to model pronunciation variation is by adding pronunciation variants to the lexicon. The problem then is how to deal with these pronunciation variants at the level of the LM.

Method 1. The easiest solution is to simply add the variants to the lexicon, and not to change the LMs at all. In this case, for every variant the probabilities for the word it belongs to are used. Since the statistics for the variants are not used, it is obvious that this is a sub-optimal solution. In the following two methods the statistics for the variants are employed.

Method 2. The second solution is to use the variants themselves (instead of the underlying words) to calculate the N-grams (Kessens et al., 1999; Schiel et al., 1998; Wester et al., 1998a). For this procedure, a transcribed corpus is needed which contains information about the realized pronunciation variants. These transcriptions can be obtained in various ways, as has been discussed in Sections 2.2 and 2.4. The goal of this method is to find the string of variants V which maximizes $P(X|V) * P(V)$.

Method 3. A third possibility is to introduce an intermediate level: $P(X|V) * P(V|W) * P(W)$. The goal now is to find the string of words W and the corresponding string of variants V that maximizes the latter equation (Cremelie and Martens, 1995, 1997, 1998, 1999; Fukada et al., 1998, 1999; Perennou and Brieussel-Pousse, 1998; Pousse and Perennou, 1997). $P(V|W)$ determines the probability of the variants given the words, while $P(W)$ describes the probabilities of sequences of words. In this case, the first probabilities can also be calculated on the basis of a transcribed corpus. However, they can also be obtained otherwise. If the pronunciation variants are generated by rule, the probabilities of these rules can be used to determine the probabilities of the pronunciation variants (Cremelie and Martens, 1998, 1999; Lehtinen and Safra, 1998). Likewise, if an ANN is used to generate pronunciation variants, the ANN itself can produce probabilities of the pronunciation variants (Deshmukh et al., 1996; Fukada et al., 1998, 1999).

It is obvious that the number of pronunciation variants is larger than the number of words. As a consequence, more parameters have to be trained for the second method than for the third. This could be a disadvantage of the second method, since sparsity of data is a common problem during the training of LMs. A way of reducing the number of parameters for both methods is to use thresholds, i.e., only pronunciation variants which occur often enough are taken into account.

Another important difference between the two methods is that in the third method the context-dependence of pronunciation variants is not modeled directly in the LM. This can be a disadvantage as pronunciation variation is often context-dependent, e.g., liaison in French (Perennou and Brieussel-Pousse, 1998; Pousse and Perennou, 1997). Within the third method, this deficiency can be overcome by using classes of words instead of the words themselves, i.e., the classes of words that do or do not allow liaison (Perennou and Brieussel-Pousse, 1998; Pousse and Perennou, 1997). The probability of a pronunciation variant for a certain class is then represented in $P(V|W)$, while the probability of sequences of word classes is stored in $P(W)$.

3. Evaluation and comparison

In the previous section, the various approaches to modeling pronunciation variation were described according to their major properties. In that presentation, the emphasis was on the various characteristics of the methods, and not so much on their merits. That is not to say that the effectiveness of a method is not important. On the contrary, the extent to which each study contributes to modeling pronunciation variation in ASR, be it by reducing the number of errors caused by pronunciation variation or by getting more insight into pronunciation variation, is a fundamental aspect, especially if we want to draw general conclusions as to the different ways in which pronunciation variation in ASR can best be addressed.

Considering that pronunciation variation research in ASR still has a long way to go, it would seem that studies that provide insight into the processes underlying pronunciation variation (e.g., Adda-Decker and Lamel, 1998, 1999; Fosler-Lussier and Morgan, 1998, 1999; Greenberg, 1998, 1999; Peters and Stubley, 1998; Strik and Cucchiarini, 1998) are very useful. However, the majority of the papers about modeling pronunciation variation for ASR focus on the effectiveness of a given method of variation modeling in improving recognition performance. It is obvious that studies of this kind could also contribute to gaining insight, if, for instance, some kind of error analysis were carried out, but this does not seem to be a priority.

The effectiveness of word-error-rate (WER) reducing studies is usually established by comparing the performance of the baseline system (the starting point) with the performance obtained after the method has been applied. For every individual study, this seems a plausible procedure. The amounts of improvement reported in the literature (see e.g., the papers in this special issue and in the proceedings of the Rolduc workshop; Strik et al., 1998) differ from almost none (and occasionally even a deterioration) to substantial ones.

In trying to draw general conclusions as to the effectiveness of these methods, one is tempted to conclude that the method for which the largest improvement was observed is the best one. In this respect some comment is in order. First, it is unlikely that there will be one single best approach, as the tasks of the various systems are very different. Second, we are not interested in finding a winner, but in understanding how pronunciation variation can best be approached. So, even a method that does not produce any significant reduction in WER may turn out to be extremely valuable because it increases our understanding of pronunciation variation. Third, it is wrong to take the change in WER as the only criterion for evaluation, because this change is dependent on at least three different factors: (1) the corpora, (2) the ASR system and (3) the baseline system. This means that improvements in WER can be compared with each other only if in the methods under study these three elements were identical or at least similar. It is obvious that in the majority of the methods presented these three elements are not kept constant, but are usually very different. In the following sections we discuss these differences and try to explain why this makes it difficult to compare the various methods and, in particular, the results obtained with each of them.

3.1. Differences between corpora

Corpora are used to gauge the performance of ASR systems. In studies on pronunciation variation modeling, many different corpora are used. The choice of a given corpus at the same time implies the choice of the task, the type of speech and the language. This means that there are at least three respects in which corpora may differ from each other.

Both with respect to task and type of speech it is possible to distinguish between cases with little pronunciation variation (carefully read speech) and cases with much more variation (conversational, spontaneous speech). Given this difference in amount of variation, it is possible that a method for pronunciation variation modeling that performs well for read speech does not perform equally well for conversational speech.

Another important aspect of the corpus is the language. Since pronunciation variation will also differ between languages, a method which gives good results in one language need not be as effective in another language. For example, Beulen et al. (1998) report improvements for English corpora while with the same method no improvements were obtained for a German corpus. Another example concerns the pronunciation variation caused by liaison in French. Perennou et al. (Perennou and Brieussel-Pousse, 1998; Pousse and Perennou, 1997) propose a method to model this type of pronunciation variation, and for their French corpus this yields an improvement. However, it remains to be seen how effective their method is in modeling pronunciation variation in languages in which other processes take place.

3.2. Differences between ASR systems

As we all know, not all ASR systems are similar. A method that is successful with one ASR system, may be less successful with another. This will already be the case for ASR systems with a similar architecture (e.g., a “standard ASR system” with the common phone-based HMMs), but it will certainly be true for ASR systems with totally different architectures. For instance, Cremelie and Martens (1998, 1999) obtain large improvements with a rule-based method for their segment-based ASR system, which does not imply that the same method will be equally successful for another type of ASR system.

Moreover, a method can be successful with a given ASR system, not so much because it models pronunciation variation in the correct way, but because it corrects for the peculiarities of the ASR system. To illustrate this point let us assume that a specific ASR system very often recognizes /n/ in certain contexts as /m/. If the method for pronunciation variation modeling replaces the proper occurrences of /n/ by /m/ in the lexicon, the performance will certainly go up. Such a transformation is likely to occur in a data-driven method in which a DP-alignment is used (see Section 2.3.).

By looking at the numbers alone (the performance before and after the method was applied), one could conclude that the method is successful. However, in this particular case, the method is successful because it corrects the errors made by the ASR system. Although one could argue that the error made by the ASR system (i.e., recognizing certain /n/s as /m/) is in fact due to pronunciation variation, the example clearly demonstrates that certain methods may work with a specific ASR system, but do not necessarily generalize to other systems.

Let us state clearly that being able to correct for the peculiarities of an ASR system is not a bad property of a method. On the contrary, if a method has this property it is almost certain that it will increase the performance of the ASR system. This is probably why in (Riley et al., 1998, 1999) it is argued that the ASR system itself should be used to make the transcriptions. The point to be made in the example above is that a posteriori it is not easy to determine which part of the improvement is due to correct modeling of pronunciation variation by the method or due to other reasons. In turn, this will make it difficult to estimate how successful a method will be for another ASR system. After all, the peculiarities of all ASR systems are not the same.

3.3. Differences in the measures used for evaluation

In the various studies, different measures are used to express the change in WER. The three measures used most often are discussed in this subsection. In order to illustrate these three measures some examples with fictitious numbers are presented in Table 1.

First, the performance is calculated for the baseline system, say $\mathrm{WER_{begin}}$. Then the method is applied, e.g., by adding pronunciation variants to the lexicon, and the performance of the new system is determined, say $\mathrm{WER_{end}}$. The absolute improvement then is

$$\%\mathrm{abs} = \mathrm{WER_{begin}} - \mathrm{WER_{end}}.$$

This can also be expressed in relative terms:

$$\%\mathrm{rel}_1 = 100 * (\mathrm{WER_{begin}} - \mathrm{WER_{end}})/\mathrm{WER_{begin}}.$$

The measure $\%\mathrm{rel}_1$ yields higher numbers than the measure %abs, but even higher numbers can be obtained by using

$$\%\mathrm{rel}_2 = 100 * (\mathrm{WER_{begin}} - \mathrm{WER_{end}})/\mathrm{WER_{end}}.$$

The last equation is generally considered to be less correct. Furthermore, for most people $\%\mathrm{rel}_1$ is more in agreement with their intuition than $\%\mathrm{rel}_2$, i.e., most people would say that an improvement from 20% to 10% WER is an improvement of 50% and not an improvement of 100% (see Table 1).

In general, $\%\mathrm{rel}_1$ is also a better measure than %abs. Most people will agree that the improvements obtained in examples 2 and 3 are more similar than those in examples 1 and 2. To summarize, $\%\mathrm{rel}_1$ is a measure that is more in line with our intuitions about the amount of improvement than %abs or $\%\mathrm{rel}_2$. Therefore, it is probably better to use the measure $\%\mathrm{rel}_1$ to express the changes in WER. In addition, $\mathrm{WER_{begin}}$ and optionally $\mathrm{WER_{end}}$ should also be specified.

3.4. Differences in the baseline systems

Another reason why it is difficult to compare methods is related to the baseline systems (the starting points) used. Let us explain why. Whatever equation is used, it is clear that the outcome of the equation depends on two numbers: $\mathrm{WER_{begin}}$ and $\mathrm{WER_{end}}$. In most studies, a lot of work is done in order to decrease $\mathrm{WER_{end}}$, and this

Table 1

Some fictitious numerical examples to illustrate the measures %abs, $\%\mathrm{rel}_1$ and $\%\mathrm{rel}_2$

Example	$\mathrm{WER_{begin}}$ (%)	$\mathrm{WER_{end}}$ (%)	%abs (%)	$\%\mathrm{rel}_1$ (%)	$\%\mathrm{rel}_2$ (%)
1	50	40	10	20	25
2	20	10	10	50	100
3	50	25	25	50	100

work is generally described in detail. However, more often than not the baseline system is not clearly described and no attempt is made to improve it. Usually, the starting point is simply an ASR system that was available at the beginning of the research, or an ASR system that is quickly trained with resources available at the beginning of the research. It is clear that for a relatively bad baseline system it is much easier to obtain improvements than for a good baseline system. For instance, a baseline system may contain errors, e.g., errors in the canonical lexicon. During research, part of these errors may be corrected, e.g., by changing the transcriptions in the lexicon. If corrections are made, similar corrections should also be made in the baseline system and $\mathrm{WER}_{\mathrm{begin}}$ should be calculated again. If this is not done, part of the resulting improvement is due to the correction of errors and possibly other sources. This makes it difficult to estimate which part of the improvement is really due to the modeling of pronunciation variation.

Besides the presence of errors, other properties of the canonical lexicon will also, to a large extent, determine the amount of improvement obtained with a certain method. Let us assume, for the sake of argument, that the canonical lexicon contains pronunciations (i.e., transcriptions) for a certain accent and speaking style (e.g., read speech). A method is then tested with a corpus that contains speech of another accent and another speaking style (e.g., conversational speech). The method succeeds in improving the lexicon in the sense that the new pronunciations in the lexicon are more appropriate for the speech in the corpus, and a large improvement in the performance is observed. Although it is clear that the method has succeeded in modeling pronunciation variation, it is also clear that the amount of improvement would have been (much) smaller if the lexicon had contained more appropriate transcriptions from the start and not those of another accent and another speech type.

In short, a large amount of research and written explanation is devoted to the reduction of $\mathrm{WER}_{\mathrm{end}}$, while relatively little effort is put in $\mathrm{WER}_{\mathrm{begin}}$. Since both quantities determine the amount of improvement, and since the baseline systems differ between studies, it becomes difficult to compare the various methods.

3.5. Objective evaluation

The question that arises at this point is: Is an objective evaluation and comparison of these methods at all possible? This question is not easy to answer. An obvious solution seems to be to use benchmark corpora and standard methods for evaluation (e.g., to give everyone the same canonical lexicon), like the NIST evaluations for automatic speech recognition and automatic speaker verification. This would solve a number of the problems mentioned above, but certainly not all of them. The most important problem that remains is the choice of the language. Like many other benchmark tests it could be (American) English. However, pronunciation variation and the ways in which it should be modeled can differ between languages, as argued above. Furthermore, for various reasons it would favor groups who do research on (American) English. Finally, using benchmarks would not solve the problem of differences between ASR systems.

Still, the large scale (D)ARPA projects and the NIST evaluations have shown that the combination of competition and objective evaluation (i.e., the possibility to obtain an objective comparison of methods) is very useful. Therefore, it seems advisable to strive towards objective evaluation methods within the field of pronunciation modeling.

4. Discussion and conclusions

One of the most common ways of modeling pronunciation variation is to add pronunciation variants to the lexicon (see Section 2.4.1). This method can be applied fairly easily and it appears to improve recognition performance. However, a problem with this approach is that certain words have numerous variants with very different frequencies of occurrence. Some quantitative data on this phenomenon can be found in Table 2 on page 50 of Greenberg (1998). For instance, if we look at the data for “that”, we can see that this word appears 328 times in the corpus used by Greenberg, that it has 117 different pronunciations and that the single most common variant only covers 11% of the pronunciations. The coverage of the other variants will probably decrease gradually from 11% to almost zero. In principle one could include all 117 variants in the lexicon and it is possible that this will improve recognition of the word "that". However, this is also likely to increase confusability. If many variants of a large number of words are included in the lexicon the confusability can increase to such an extent that recognition performance may eventually decrease. This implies that variant selection constitutes an essential part of this approach.

An obvious criterion for variant selection is frequency of occurrence. Adding very frequent variants is likely to produce a more substantial improvement than adding infrequent variants. However, there is no clear distinction between frequent and infrequent pronunciation variants. Furthermore, besides frequency of occurrence, there will be other important factors that influence recognition performance. For instance, some pronunciation variants will probably constitute no problem for the ASR system, in the sense that they will be recognized correctly even though they (slightly) differ from the representation stored in the lexicon. Other spoken variants will probably cause frequent errors during recognition. In order to improve the performance of the ASR system, it is necessary to know which variants cause recognition errors (and which do not). Furthermore, adding pronunciation variants to the lexicon can solve some recognition errors, but it will certainly also introduce new ones. To optimize performance one should add those variants for which the balance between solving old errors and introducing new errors is positive. Confusability during the decoding process is a central issue in this respect. However, it will be difficult to predict a priori what the confusability during decoding will be. A manner in which our insight on this topic could be enhanced, is by doing error analysis, as will be discussed below.

In most studies mentioned above the emphasis was on reduction of the error rates. However, the difference in the error rates of two systems is only a global measure which does not provide information about the details of the differences in the recognition results. Consequently, in most studies it is not possible to find out how and why improvements were obtained. In order to do so the recognition errors should be studied in more detail, i.e., more detailed error analysis should be carried out. This can be done by comparing the errors in the recognition results between the old system and the new one. In addition, error analysis could be used not only post hoc, to test the effect of a specific method, but also before applying the method. For instance, it would be informative to know beforehand how many and what kind of errors are made so as to be able to estimate the maximum amount of improvement that can be achieved. In turn this could constitute a criterion in deciding whether to test the method at all.

One of the reasons why error analysis is often omitted is related to the availability of data. In order to test the performance of an ASR system a test corpus is needed. According to the rules of the game, nothing must be known about the test corpus. As soon as you start doing a detailed error analysis on the test corpus you learn about the corpus. In turn this entails that this specific corpus can no longer be used for objective evaluation. For instance, suppose that error analysis revealed that recognition of the word "that" is problematic. In this case one could try to solve this problem, thus improving the performance of the ASR system on that specific corpus. Quite clearly, this is not fair. A possible alternative would be to use the training corpus for error analysis. Since in this case the material used for training and for testing is the same, this could influence the outcome of the error analysis. The best option probably would be to use an independent development test set for error analysis.

At this point, it may be useful to try to make a general assessment of the state of the art in research on modeling pronunciation variation for ASR. For example, we could start by relating the results obtained so far to the expectations researchers had at the beginning. It is difficult, though, to estimate the researchers' expectations about the gain in recognition performance that could be obtained by modeling pronunciation variation. In any case, judging by the effort that has gone in this type of research one could conclude that there were high expectations. The results reported so far vary from 0% to 20% relative reduction in the WER. These findings can be interpreted either positively or negatively. Positively: modeling pronunciation variation often improves recognition performance, sometimes even by 20%. Negatively: sometimes recognition performance increases by about 20%, but in most cases improvements are marginal. At the Rolduc workshop the general feeling seemed to be that the results obtained so far did not live up to the expectations.

However, the general idea also seemed to be that research on pronunciation variation modeling has made different useful contributions to ASR. For example, this research has shown the importance of systematic lexicon design and has produced improved, more consistent lexica. Furthermore, different methods for pronunciation variation modeling have been proposed that, in general, lead to some improvement. For instance, it has now become standard procedure to include multiple variants for some of the words in the lexicon. Finally, this research has produced the knowledge that the problem of pronunciation variation is not a simple one.

In any case it is clear that the right solutions have not been found yet. On the contrary, the modest improvements suggest that we are just at the beginning of the path towards the optimal solutions. In other words, more research should be carried out, but the questions are: what kind of research, in which direction?

First of all, more fundamental research is needed to gather more knowledge on pronunciation variation. The papers at the Rolduc workshop that mainly aimed at gaining insight into pronunciation variation are (Adda-Decker and Lamel, 1998, 1999; Fosler-Lussier and Morgan, 1998, 1999; Greenberg, 1998, 1999; Peters and Stubley, 1998; Strik and Cucchiarini, 1998). In (Adda-Decker and Lamel, 1998, 1999; Fosler-Lussier and Morgan, 1998, 1999; Greenberg, 1998, 1999) various frequency counts on the basis of corpora are reported, in (Peters and Stubley, 1998) a method is presented for visualizing speech trajectories, and in (Strik and Cucchiarini, 1998) an overview of the literature is presented. In addition to analysis of speech corpora, it would be useful to study the speech production processes that lead to pronunciation variation and the type of problems that pronunciation variation causes in human speech perception and in ASR.

Finally, it is worth mentioning that at present most researchers use "standard ASR systems" based on discrete segmental representations, HMMs to model the segments, and features that are computed per frame (usually cepstral features and their derivatives). Possibly, the underlying assumptions in these standard ASR systems are not optimal. One of the assumptions is that speech is made up of discrete segments, usually phone(me)s. Although this has long been one of the assumptions in linguistics too, the idea that speech can be phonologically represented as a sequence of discrete entities (the "absolute slicing hypothesis", as formulated in (Goldsmith, 1976, pp. 16–17)) has proved to be untenable. In non-linear, autosegmental phonology (Goldsmith, 1976, 1990) an analysis has been proposed in which different features are placed on different tiers. The various tiers represent the parallel activities of the articulators in speech, which do not necessarily begin and end simultaneously. In turn the tiers are connected by association lines. In this way, it is possible to indicate that the mapping between tiers is not always one to one. Assimilation phenomena can then be represented by the spreading of one feature from one segment to the adjacent one. On the basis of this theory, Li Deng and his colleagues have built ASR systems with which promising results have been obtained (Deng and Sun, 1994).

Another important assumption of standard ASR systems, which is of course related to the first one, is that feature values can be calculated locally, per individual frame. In addition to this knowledge on the static properties of features, information on their dynamic characteristics can be obtained by computing derivatives of these features. However, the problem remains that the analysis window on which feature values are calculated is very small, while it is known from research on human perception that for perceiving one speech sound subjects rely on information contained in adjacent sounds. If this works satisfactorily in human perception, it is perhaps worthwhile to investigate whether there are better feature representations for ASR, which go beyond segment boundaries and span a larger window. Since the standard ASR systems described above have given promising results so far, researchers are reluctant to abandon this path. The future will show whether this is the right path or a dead end.

Acknowledgements

The research of Dr. H. Strik has been made possible by a fellowship of the Royal Netherlands Academy of Arts and Sciences. We would like to thank Loe Boves, Judith Kessens, Mirjam Wester, Febe de Wet and three anonymous reviewers for their comments on previous versions of this article.

References

Adda-Decker, M., Lamel, L., 1998. Pronunciation variants across systems, languages and speaking style. In: Strik, H., Kessens, J.M., Wester, M. (Eds.), Proceedings of the ESCA Workshop on Modeling Pronunciation Variation for Automatic Speech Recognition, Rolduc, Kerkrade, 4–6 May 1998. A^{2}RT, University of Nijmegen, pp. 1–6.

Adda-Decker, M., Lamel, L., 1999. Pronunciation variants across system configuration, language and speaking style. Speech Communication 29 (2–4), 83–98.

Aubert, X., Dugast, C., 1995. Improved acoustic–phonetic modeling in Philips' dictation system by handling liaisons and multiple pronunciations. In: Proceedings of Eurospeech-95, Madrid, pp. 767–770.

Bacchiani, M., Ostendorf, M., 1998. Joint acoustic unit design and lexicon generation. In: Strik, H., Kessens, J.M., Wester, M. (Eds.), Proceedings of the ESCA Workshop on Modeling Pronunciation Variation for Automatic Speech Recognition, Rolduc, Kerkrade, 4–6 May 1998. A^{2}RT, University of Nijmegen, pp. 7–12.

Bacchiani, M., Ostendorf, M., 1999. Joint lexicon, acoustic unit inventory and model design. Speech Communication 29 (2–4), 99–114.

Barnett, J., 1974. A phonological rule compiler. In: Erman., L. (Ed.), Proceedings of the IEEE Symposium on Speech Recognition, Carnegie-Mellon University (IEEE Catalog No. 74CH0878-9 AE), 15–19 April, Pittsburgh, PA, pp. 188–192.

Bell, A., 1984. Language style as audience design. Language in Society 13 (2), 145–204.

Beulen, K., Ortmanns, S., Eiden, A., Martin, S., Welling, L., Overmann, J., Ney, H., 1998. Pronunciation modelling in the RWTH large vocabulary speech recognizer. In: Strik, H., Kessens, J.M., Wester, M. (Eds.), Proceedings of the ESCA Workshop on Modeling Pronunciation Variation for Automatic Speech Recognition, Rolduc, Kerkrade, 4–6 May 1998. A^{2}RT, University of Nijmegen, pp. 13–16.

Blackburn, C.S., Young, S.J., 1995. Towards improved speech recognition using a speech production model. In: Proceedings of EuroSpeech-95, Madrid, pp. 1623–1626.

Blackburn, C.S., Young, S.J., 1996. Pseudo-articulatory speech synthesis for recognition using automatic feature extraction from X-ray data. In: Proceedings of ICSLP-96, Philadelphia, pp. 969–972.

Bonaventura, P., Gallocchio, F., Mari, J., Micca, G., 1998. Speech recognition methods for non-native pronunciation variations. In: Strik, H., Kessens, J.M., Wester, M. (Eds.), Proceedings of the ESCA Workshop on Modeling Pronunciation Variation for Automatic Speech Recognition, Rolduc, Kerkrade, 4–6 May 1998. A^{2}RT, University of Nijmegen, pp. 17–22.

Cohen, M., 1989. Phonological structures for speech recognition. Ph.D. Thesis, University of California, Berkeley, USA.

Cohen, P.S., Mercer, R.L., 1974. The phonological component of an automatic speech-recognition system. In: Erman, L. (Ed.), Proceedings of the IEEE Symposium on Speech Recognition, Carnegie-Mellon University (IEEE Catalog No. 74CH0878-9 AE), 15–19 April, Pittsburgh, PA, pp. 177–187.

Cohen, P.S., Mercer, R.L., 1975. The phonological component of an automatic speech-recognition system. In: Reddy, D.R. (Ed.), Speech Recognition. Academic Press, New York, 1975, pp. 275–320.

Coupland, N., 1984. Accommodation at work: Some phonological data and their implications. International Journal of the Sociology of Language 46, 49–70.

Cremelie, N., Martens, J.-P., 1995. On the use of pronunciation rules for improved word recognition. In: Proceedings of Eurospeech-95, Madrid, pp. 1747–1750.

Cremelie, N., Martens, J.-P., 1997. Automatic rule-based generation of word pronunciation networks. In: Proceedings of EuroSpeech-97, Rhodes, pp. 2459–2462.

Cremelie, N., Martens, J.-P., 1998. In search of pronunciation rules. In: Strik, H., Kessens, J.M., Wester, M. (Eds.), Proceedings of the ESCA Workshop on Modeling Pronunciation Variation for Automatic Speech Recognition, Rolduc, Kerkrade, 4–6 May 1998. A^{2}RT, University of Nijmegen, pp. 23–28.

Cremelie, N., Martens, J.-P., 1999. In search of better pronunciation models for speech recognition. Speech Communication 29 (2–4), 115–136.

Deng, L., Sun, D., 1994. A statistical approach to automatic speech recognition using the atomic speech units constructed from overlapping articulatory features. Journal of the Acoustical Society of America 95 (5), 2702–2719.

Deshmukh, N., Weber, M., Picone, J., 1996. Automated generation of *N*-best pronunciations of proper nouns. In: Proceedings of ICASSP-96, Atlanta, pp. 283–286.

Downey, S., Wiseman, R., 1997. Dynamic and static improvements to lexical baseforms. In: Proceedings of Eurospeech-97, Rhodes, pp. 1027–1030.

Erman, L., 1974. Proceedings of the IEEE Symposium on Speech Recognition, Carnegie-Mellon University, (IEEE Catalog No. 74CH0878-9 AE) 15–19 April 1974, Pittsburgh, PA, 295 pp.

Eskenazi, M., 1993. Trends in speaking styles research. In: Proceedings of Eurospeech-93, Berlin, pp. 501–509.

Ferreiros, J., Macías-Guarasa, J., Pardo, J.M., Villarrubia, L., 1998. Introducing multiple pronunciations in Spanish speech recognition systems. In: Strik, H., Kessens, J.M., Wester, M. (Eds.), Proceedings of the ESCA Workshop on Modeling Pronunciation Variation for Automatic Speech Recognition, Rolduc, Kerkrade, 4–6 May 1998. A^{2}RT, University of Nijmegen, pp. 29–34.

Finke, M., Waibel, A., 1997. Speaking mode dependent pronunciation modeling in large vocabulary conversational speech recognition. In: Proceedings of EuroSpeech-97, Rhodes, pp. 2379–2382.

Flach, G., 1995. Modelling pronunciation variability for spectral domains. In: Proceedings of Eurospeech-95, Madrid, pp. 1743–1746.

Fosler-Lussier, E., Morgan, N., 1998. Effects of speaking rate and word frequency on conversational pronunciations. In: Strik, H., Kessens, J.M., Wester, M. (Eds.), Proceedings of the ESCA Workshop on Modeling Pronunciation Variation for Automatic Speech Recognition, Rolduc, Kerkrade, 4–6 May 1998. A^{2}RT, University of Nijmegen, pp. 35–40.

Fosler-Lussier, E., Morgan, N., 1999. Effects of speaking rate and word frequency on pronunciations in conversational speech. Speech Communication 29 (2–4) 137–158.

Friedman, J., 1974. Computer exploration of fast speech rules. In: Erman, L. (Ed.), Proceedings of the IEEE Symposium on Speech Recognition, Carnegie-Mellon University (IEEE Catalog No. 74CH0878-9 AE), 15–19 April 1974, Pittsburgh, PA, pp. 197–203.

Fukada, T., Sagisaka, Y., 1997. Automatic generation of a pronunciation dictionary based on a pronunciation network. In: Proceedings of EuroSpeech-97, Rhodes, pp. 2471–2474.

Fukada, T., Yoshimura, T., Sagisaka, Y., 1998. Automatic generation of multiple pronunciations based on neural networks and language statistics. In: Strik, H., Kessens, J.M., Wester, M. (Eds.), Proceedings of the ESCA Workshop on Modeling Pronunciation Variation for Automatic Speech Recognition, Rolduc, Kerkrade, 4–6 May 1998. A^{2}RT, University of Nijmegen, pp. 41–46.

Fukada, T., Yoshimura, T., Sagisaka, Y., 1999. Automatic generation of multiple pronunciations based on neural networks. Speech Communication 27 (1), 63–73.

Giles, H., Powesland, P., 1975. Speech Style and Social Evaluation. Cambridge University Press, Cambridge.

Giles, H., Smith, P., 1979. Accommodation theory: Optimal levels of convergence. In: Giles, H., stClair, R. (Eds.), Language and Social Psychology, Blackwell, Oxford.

Godfrey, J.J., Ganapathiraju, A., Ramalingam, C.S., Picone, J., 1997. Microsegment-based connected digit recognition. In: Proceedings of ICASSP-97, Munich, pp. 1755–1758.

Goldsmith, J., 1976. Autosegmental phonology. Doctoral thesis, Massachussets Institute of Technology, Cambridge. Indiana University Linguistics Club, Bloomington, Indiana; Garland, New York, 1979.

Goldsmith, J.A., 1990. Autosegmental and Metrical Phonology. Blackwell, Oxford.

Greenberg, S., 1998. Speaking in shorthand – A syllable-centric perspective for understanding pronunciation variation. In: Strik, H., Kessens, J.M., Wester, M. (Eds.), Proceedings of the ESCA Workshop on Modeling Pronunciation Variation for Automatic Speech Recognition, Rolduc, Kerkrade, 4–6 May 1998. A^{2}RT, University of Nijmegen, pp. 47–56.

Greenberg, S., 1999. Speaking in shorthand – A syllable-centric perspective for understanding pronunciation variation. Speech Communication 29 (2–4) 159–176.

Heine, H., Evermann, G., Jost, U., 1998. An HMM-based probabilistic lexicon. In: Strik, H., Kessens, J.M., Wester, M. (Eds.), Proceedings of the ESCA Workshop on Modeling Pronunciation Variation for Automatic Speech Recognition, Rolduc, Kerkrade, 4–6 May 1998. A^{2}RT, University of Nijmegen, pp. 57–62.

Holmes, W.J., Russell, M.J., 1996. Modeling speech variability with segmental HMMs. In: Proceedings of ICASSP-96, Atlanta, pp. 447–450.

Holter, T., 1997. Maximum likelihood modelling of pronunciation in automatic speech recognition. Ph.D. Thesis, Norwegian University of Science and Technology, December 1997.

Holter, T., Svendsen, T., 1998. Maximum likelihood modelling of pronunciation variation. In: Strik, H., Kessens, J.M., Wester, M. (Eds.), Proceedings of the ESCA Workshop on Modeling Pronunciation Variation for Automatic Speech Recognition, Rolduc, Kerkrade, 4–6 May 1998. A^{2}RT, University of Nijmegen, pp. 63–66.

Holter, T., Svendsen, T., 1999. Maximum likelihood modelling of pronunciation variation. Speech Communication 29 (2–4) 177–191.

Imai, T., Ando, A., Miyasaka, E., 1995. A new method for automatic generation of speaker-dependent phonological rules. In: Proceedings of ICASSP-95, Detroit, pp. 864–867.

Jelinek, F., Bahl, L.R., Mercer, R.L., 1974. Design of a linguistic statistical decoder for the recognition of continuous speech. In: Erman, L. (Ed.), Proceedings of the IEEE Symposium on Speech Recognition, Carnegie-Mellon University (IEEE Catalog No. 74CH0878-9 AE), 15–19 April 1974, Pittsburgh, PA, pp. 255–260.

Kaisse, E., 1985. Connected Speech: The Interaction of Syntax and Phonology. Academic Press, Orlando.

Kessens, J., Wester, M., 1997. Improving recognition performance by modelling pronunciation variation. In: Proceedings of the CLS opening Academic Year '97–'98, pp. 1–20 (http://lands.let.kun.nl/literature/kessens.1997.1.html).

Kessens, J.M., Wester, M., Strik, H., 1999. Improving the performance of a Dutch CSR by modelling within-word and cross-word pronunciation variation. Speech Communication 29 (2–4) 193–207.

Kipp, A., Wesenick, M.-B., Schiel, F., 1996. Automatic detection and segmentation of pronunciation variants in German speech corpora. In: Proceedings of ICSLP-96, Philadelphia, pp. 106–109.

Kipp, A., Wesenick, M.-B., Schiel, F., 1997. Pronunciation modeling applied to automatic segmentation of spontaneous speech. In: Proceedings of EuroSpeech-97, Rhodes, pp. 1023–1026.

Labov, W., 1972. Sociolinguistic Patterns. University of Pennsylvania Press, Philadelphia.

Lamel, L., Adda, G., 1996. On designing pronunciation lexicons for large vocabulary continuous speech recognition. In: Proceedings of ICSLP-96, Philadelphia, pp. 6–9.

Laver, J., 1994. Principles of Phonetics. Cambridge University Press, Cambridge.

Lehtinen, G., Safra, S., 1998. Generation and selection of pronunciation variants for a flexible word recognizer. In: Strik, H., Kessens, J.M., Wester, M. (Eds.), Proceedings of the ESCA Workshop on Modeling Pronunciation Variation for Automatic Speech Recognition, Rolduc, Kerkrade, 4–6 May 1998. A^{2}RT, University of Nijmegen, pp. 67–72.

Mercer, R., Cohen, P., 1987. A method for efficient storage and rapid application of context-sensitive phonological rules for automatic speech recognition. IBM J. Res. Develop. 31 (1), 81–90.

Mirghafori, N., Fosler, E., Morgan, N., 1995. Fast speakers in large vocabulary continuous speech recognition: analysis and antidotes. In: Proceedings of EuroSpeech-95, Madrid, pp. 491–494.

Mokbel, H., Jouvet, D., 1998. Derivation of the optimal phonetic transcription set for a word from its acoustic realisations. In: Strik, H., Kessens, J.M., Wester, M. (Eds.), Proceedings of the ESCA Workshop on Modeling Pronunciation Variation for Automatic Speech Recognition, Rolduc, Kerkrade, 4–6 May 1998. A^{2}RT, University of Nijmegen, pp. 73–78.

Mouria-Beji, F., 1998. Context and speed dependent phonemic models for continuous speech recognition. In: Strik, H., Kessens, J.M., Wester, M. (Eds.), Proceedings of the ESCA Workshop on Modeling Pronunciation Variation for Automatic Speech Recognition, Rolduc, Kerkrade, 4–6 May 1998. A^{2}RT, University of Nijmegen, pp. 79–84.

Murray, I.R., Arnott, J.L., 1993. Toward the simulation of emotion in synthetic speech: A review of the literature on human vocal emotion. Journal of the Acoustical Society of America 93 (2), 1097–1108.

Nock, H.J., Young, S.J., 1998. Detecting and correcting poor pronunciations for multiword units. In: Strik, H., Kessens, J.M., Wester, M. (Eds.), Proceedings of the ESCA Workshop on Modeling Pronunciation Variation for Automatic Speech Recognition, Rolduc, Kerkrade, 4–6 May 1998. A^{2}RT, University of Nijmegen, pp. 85–90.

O'Malley, M.H., Cole, A., 1974. Testing phonological rules. In: Erman, L. (Ed.), Proceedings of the IEEE Symposium on Speech Recognition, Carnegie-Mellon University (IEEE Catalog No. 74CH0878-9 AE), 15–19 April 1974, Pittsburgh, PA, pp. 193–196.

Oshika, B.T., Zue, V.W., Weeks, R.V., Neu, H., 1974. The role of phonological rules in speech understanding research. In: Erman, L. (Ed.), Proceedings of the IEEE Symposium on Speech Recognition, Carnegie-Mellon University (IEEE Catalog No. 74CH0878-9 AE), 15–19 April 1974, Pittsburgh, PA pp. 204–207.

Perennou, G., Brieussel-Pousse, L., 1998. Phonological component in automatic speech recognition. In: Strik, H., Kessens, J.M., Wester, M. (Eds.), Proceedings of the ESCA Workshop on Modeling Pronunciation Variation for Automatic Speech Recognition, Rolduc, Kerkrade, 4–6 May 1998. A^{2}RT, University of Nijmegen, pp. 91–96.

Peters, S.D., Stubley, P., 1998. Visualizing speech trajectories. In: Strik, H., Kessens, J.M., Wester, M. (Eds.), Proceedings of the ESCA Workshop on Modeling Pronunciation Variation for Automatic Speech Recognition, Rolduc, Kerkrade, 4–6 May 1998. A^{2}RT, University of Nijmegen, pp. 97–102.

Polzin, T.S., Waibel, A.H., 1998. Pronunciation variations in emotional speech. In: Strik, H., Kessens, J.M., Wester, M. (Eds.), Proceedings of the ESCA Workshop on Modeling Pronunciation Variation for Automatic Speech Recognition, Rolduc, Kerkrade, 4–6 May 1998. A^{2}RT, University of Nijmegen, pp. 103–108.

Pousse, L., Perennou, G., 1997. Dealing with pronunciation variants at the language model level for automatic continuous speech recognition of French. In: Proceedings of Eurospeech-97, Rhodes, pp. 2727–2730.

Rabinowitz, A.S., 1974. Phonetic to graphemic transformation by use of a stack procedure. In: Erman, L. (Ed.), Proceedings of the IEEE Symposium on Speech Recognition, Carnegie-Mellon University (IEEE Catalog No. 74CH0878-9 AE), 15–19 April 1974, Pittsburgh, PA, pp. 212–217.

Ravishankar, M., Eskenazi, M., 1997. Automatic generation of context-dependent pronunciations. In: Proceedings of EuroSpeech-97, Rhodes, pp. 2467–2470.

Riley, M., Byrne, W., Finke, M., Khudanpur, S., Ljolje, A., McDonough, J., Nock, H., Saraclar, M., Wooters, C., Zavaliagkos, G., 1998. Stochastic pronunciation modelling from hand-labelled phonetic corpora. In: Strik, H., Kessens, J.M., Wester, M. (Eds.), Proceedings of the ESCA Workshop on Modeling Pronunciation Variation for Automatic Speech Recognition, Rolduc, Kerkrade, 4–6 May 1998. A^{2}RT, University of Nijmegen, pp. 109–116.

Riley, M., Byrne, W., Finke, M., Khudanpur, S., Ljolje, A., McDonough, J., Nock, H., Saraclar, M., Wooters, C., Zavaliagkos, G., 1999. Stochastic pronunciation modelling from hand-labelled phonetic corpora. Speech Communication 29 (2–4) 209–224.

Ristad, E.S., Yianilos, P.N., 1998. A surficial pronunciation model. In: Strik, H., Kessens, J.M., Wester, M. (Eds.), Proceedings of the ESCA Workshop on Modeling Pronunciation Variation for Automatic Speech Recognition, Rolduc, Kerkrade, 4–6 May 1998. A^{2}RT, University of Nijmegen, pp. 117–120.

Roach, P., Arnfield, S., 1998. Variation information in pronunciation dictionaries. In: Strik, H., Kessens, J.M., Wester, M. (Eds.), Proceedings of the ESCA Workshop on Modeling Pronunciation Variation for Automatic Speech Recognition, Rolduc, Kerkrade, 4–6 May 1998. A^{2}RT, University of Nijmegen, pp. 121–124.

Roe, D.B., Riley, M.D., 1994. Prediction of word confusabilities for speech recognition. In: Proceedings of ICSLP-94, Yokohama, pp. 227–230.

Romaine, S., 1980. Stylistic variation and evaluative reactions to speech. Language and Speech 23, 213–232.

Rovner, P., Makhoul, J., Wolf, J., Colarusso, J., 1974. Where the words are: lexical retrieval in a speech understanding system. In: Erman, L. (Ed.), Proceedings of the IEEE Symposium on Speech Recognition, Carnegie-Mellon University (IEEE Catalog No. 74CH0878-9 AE), 15–19 April 1974, Pittsburgh, PA, pp. 160–164.

Safra, S., Lehtinen, G., Huber, K., 1998. Modeling pronunciation variations and coarticulation with finite-state transducers in CSR. In: Strik, H., Kessens, J.M., Wester, M. (Eds.), Proceedings of the ESCA Workshop on Modeling Pronunciation Variation for Automatic Speech Recognition, Rolduc, Kerkrade, 4–6 May 1998. A^{2}RT, University of Nijmegen, pp. 125–130.

Scherer, K.R., Giles, H., 1979. Social Markers in Speech. Cambridge University Press, Cambridge.

Schiel, F., Kipp, A., Tillmann, H.G., 1998. Statistical modelling of pronunciation: It's not the model, it's the data. In: Strik, H., Kessens, J.M., Wester, M. (Eds.), Proceedings of the ESCA Workshop on Modeling Pronunciation Variation for Automatic Speech Recognition, Rolduc, Kerkrade, 4–6 May 1998. A^{2}RT, University of Nijmegen, pp. 131–136.

Shockey, L., Erman, L.D., 1974. Sub-lexical levels in the HEARSAY II speech understanding system. In: Erman, L. (Ed.), Proceedings of the IEEE Symposium on Speech Recognition, Carnegie-Mellon University (IEEE Catalog No. 74CH0878-9 AE), 15–19 April 1974, Pittsburgh, PA, pp. 208–210.

Sloboda, T., Waibel, A., 1996. Dictionary learning for spontaneous speech recognition. In: Proceedings of ICSLP-96, Philadelphia, pp. 2328–2331.

Strik, H., 1998. Publications on pronunciation variation and ASR. http://lands.let.kun.nl/TSpublic/strik/pron-var/references.html.

Strik, H., Cucchiarini, C., 1998. Modeling pronunciation variation for ASR: overview and comparison of methods. In: Strik, H., Kessens, J.M., Wester, M. (Eds.), Proceedings of the ESCA Workshop on Modeling Pronunciation Variation for Automatic Speech Recognition, Rolduc, Kerkrade, 4–6 May 1998. A^{2}RT, University of Nijmegen, pp. 137–144.

Strik, H., Kessens, J.M., Wester, M., 1998. Proceedings of the ESCA Workshop on Modeling Pronunciation Variation for Automatic Speech Recognition, Rolduc, Kerkrade, 4–6 May 1998. A^{2}RT, University of Nijmegen, 168 pp.

Svendsen, T., Soong, F., Purnhagen, H., 1995. Optimizing acoustic baseforms for HMM-based speech recognition. In: Proceedings of EuroSpeech-95, Madrid, pp. 783–786.

Tappert, C.C., 1974. Experiments with a tree search method for converting noisy phonetic representation into standard orthography. In: Erman, L. (Ed.), Proceedings of the IEEE Symposium on Speech Recognition, Carnegie-Mellon University (IEEE Catalog No. 74CH0878-9 AE), 15–19 April 1974, Pittsburgh, PA, pp. 261–266.

Torre, D., Villarrubia, L., Hernández, L., Elvira, J.M., 1997. Automatic alternative transcription generation and vocabulary selection for flexible word recognizers. In: Proceedings of ICASSP-97, Munich, pp. 1463–1466.

Wesenick, M.-B., 1996. Automatic generation of German pronunciation variants. In: Proceedings of ICSLP-96, Philadelphia, pp. 125–128.

Wester, M., Kessens, J.M., Strik, H., 1998a. Improving the performance of a Dutch CSR by modelling pronunciation variation. In: Strik, H., Kessens, J.M., Wester, M. (Eds.), Proceedings of the ESCA Workshop on Modeling Pronunciation Variation for Automatic Speech Recognition, Rolduc, Kerkrade, 4–6 May 1998. A^{2}RT, University of Nijmegen, pp. 145–150.

Wester, M., Kessens, J.M., Cucchiarini, C., Strik, H., 1998b. Selection of pronunciation variants in spontaneous speech: Comparing the performance of man and machine. In: Proceedings of the ESCA workshop, SPoSS 98 – Sound Patterns of Spontaneous Speech: Production and Perception: Aix-en-Provence, France, 24–26 September 1998, pp. 157–160.

Wester, M., Kessens, J.M., Cucchiarini, C., Strik, H., 1999. Comparison between expert listeners and continuous speech recognizers in selecting pronunciation variants. In: Proceedings of the 14th International Congress of Phonetic Sciences (ICPhS-99), San Fransico, USA, 1999.

Williams, G., Renals, S., 1998. Confidence measures for evaluating pronunciation models. In: Strik, H., Kessens, J.M., Wester, M. (Eds.), Proceedings of the ESCA Workshop on Modeling Pronunciation Variation for Automatic Speech Recognition, Rolduc, Kerkrade, 4–6 May 1998. A^{2}RT, University of Nijmegen, pp. 151–156.

Wiseman, R., Downey, S., 1998. Dynamic and static improvements to lexical baseforms. In: Strik, H., Kessens, J.M., Wester, M. (Eds.), Proceedings of the ESCA Workshop on Modeling Pronunciation Variation for Automatic Speech Recognition, Rolduc, Kerkrade, 4–6 May 1998. A^{2}RT, University of Nijmegen, pp. 157–162.

Zeppenfeld, T., Finke, M., Ries, K., Westphal, M., Waibel, A., 1997. Recognition of conversational speech using the JANUS speech engine. In: Proceedings of ICASSP-97, Munich, pp. 1815–1818.