Pronunciation adaptation at the lexical level

Helmer Strik

A2RT, Dept. of Language and Speech,
Univ. of Nijmegen, the Netherlands
Strik@let.kun.nl
http://lands.let.kun.nl/TSpublic/strik/

Abstract

There are various kinds of adaptation which can be used to enhance the performance of automatic speech recognizers. This paper is about pronunciation adaptation at the lexical level, i.e. about modeling pronunciation variation at the lexical level. In the early years of automatic speech recognition (ASR) research, the amount of pronunciation variation was limited by using isolated words. Since the focus gradually shifted from isolated words to conversational speech, the amount of pronunciation variation present in the speech signals has increased, as has the need to model it. This is reflected by the growing attention for this topic. In this paper, an overview of the studies on lexicon adaptation is presented. Furthermore, many examples are mentioned of situations in which lexicon adaptation is likely to improve the performance of speech recognizers. Finally, it is argued that some assumptions made in current standard ASR systems are not in line with the properties of the speech signals. Consequently, the problem of pronunciation variation at the lexical level probably cannot be solved by simply adding new transcriptions to the lexicon, as it is generally done at the moment.

1. Introduction

> *"...Yet of those which do survive, the best adapted individuals, supposing that there is any variability in a favourable direction, will tend to propagate their kind in larger numbers than the less well adapted."* Charles Darwin, 1859 [18: p. 93]

During human evolution, the vocal organs adapted themselves in such a way that producing speech sounds became possible (which was not the original function of the vocal organs). Simultaneously, our perceptive system adapted itself in order to be able to process those speech sounds. Nowadays, we are trying to adapt automatic speech recognizers in order to improve their processing of those speech sounds that humans learned to produce and understand throughout a long period of evolution.

The theme of this workshop is adaptation for automatic speech recognition. Various kinds of adaptation are possible. In the announcement of this workshop the following types were mentioned: speaker adaptation, lexicon/pronunciation adaptation, language model adaptation, database/environment adaptation, noise/channel compensation. This paper will focus on lexicon/pronunciation adaptation. For a discussion of the other kinds of adaptation to reader is referred to the other papers in this proceedings.

Before I explain what lexicon/pronunciation adaptation stands for in this paper (at the end of the current paragraph), I would like to make a few remarks about the lexicon first. Whereas acoustic models and language models are generally the output of an optimization procedure, this is not the case for the lexicon. The lexicon, together with a corpus, is usually the input, and not the output, of a training procedure (exceptions are the procedures described in [4, 42, 43, 85]). Furthermore, the lexicon is the interface between the words and the acoustics. The lexicon defines the acoustic-phonetic units used during recognition, which are usually phones (although other units have been studied, see e.g. [4, 19, 34, 42, 43, 85]). The pronunciations present in the lexicon are transcriptions in terms of these acoustic-phonetic units. The lexicon can be adapted by adding new words to the lexicon, in order to reduce the out-of-vocabulary (OOV) rate. This will certainly lower the word error rate. However, this type of lexicon adaptation will not be addressed in this paper. In this paper I will deal with the kind of lexicon adaptation that is necessary to model pronunciation variation, i.e. pronunciation adaptation at the lexical level.

The need for modeling pronunciation variation in ASR originates from the simple fact that the words of a language are pronounced in many different ways as a result of variations in speaking style [24], degree of formality [53, 55], interlocutor [15, 32, 33], environment [46, 47], speech disability [22, 52, 58, 59, 89], accent or dialect [55], socioeconomic factors [82], anatomical differences, and emotional status [63, 69]. In the early years of ASR research, the amount of pronunciation variation was limited by using isolated words. Isolated word recognition requires speakers 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 definitely unnatural. Therefore, attempts were made in ASR research to improve technology, so that it could handle less artificial speech. As a result, the type of speech used in ASR research has gradually progressed from isolated words, through connected words and carefully read speech, to conversational or spontaneous 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 current systems.

Pronunciation adaptation at the lexical level has been a research topic since the early 1970s. For instance, many articles in the proceedings of 'the IEEE Symposium on Speech Recognition' from April 1974 [23] mention the need to include multiple pronunciations in the lexicon and suggest using phonological rules to generate these variants [5, 13, 30, 45, 65, 66, 71, 78, 84, 90]. Lately, there has been an increase in the amount of re-

search on this topic (see e.g. the many references below). An overview of the approaches described in the literature is presented in section 3, followed by a discussion in section 4. But first some examples of lexicon adaptation are given in section 2.

2. Examples of lexicon adaptation

In this section, a number of examples of lexicon adaptation are presented. These examples will illustrate that adaptation can be used in various situations, and can be done in various ways depending on the material that is available, on the target to which the system should be adapted, and on other factors that are described below.

Adaptation can be done on-line, while a user is speaking to the speech recognizer, or off-line, during research and development. The off-line adaptation can be done manually (by an expert) or automatically, whereas the on-line adaptation should always be done automatically (i.e. it cannot be done by the user). In general, the goal of off-line adaptation is to make the ASR system more user independent (in the sense that the speech recognizer should be more capable of recognizing many different speakers), whereas the goal of on-line adaptation is to make the speech recognizer more user dependent (so that it better recognizes this particular user).

The kind of adaptation that can be performed depends on the material that is available for adaptation, both the kind and the amount of material. The amount of available material can differ considerably across situations. For off-line adaptation existing corpora are used, or new corpora are collected, and the amount of material is delimited by the size of these corpora. During on-line adaptation the amount of material depends on the application. For instance, some spoken dialogue systems are used only for a short time by a single user (just a couple of questions and answers), whereas dictation systems and voice activated mobile telephones are generally used by the same speaker for a much longer period. The kind of material available for lexicon adaptation can be speech signals, text or both. In order to derive entries for the lexicon from speech signals and text, transcription and grapheme-to-phoneme conversion are needed, respectively. It is obvious that text can only be used if the interface allows entering text, as is the case for dictation systems and voice activated mobile telephones. In all other cases, adaptation can only be done on the basis of speech signals.

Another important factor is the target to which the system should be adapted. The target can be a single user, an accent, a dialect, or a particular language background. The latter is the case when a non-native speaks to a system, not using his mother tongue, but the language of the system. Computer-assisted-language-learning (CALL) systems are specially developed for non-native users, and it is now becoming customary to integrate ASR components into these CALL systems (see e.g. [25], and many papers in the proceedings of the ESCA workshop STiLL: Speech Technology in Language Learning [11]). However, even speech recognizers that are not specially developed for non-native users can and will be used by non-native speakers, e.g. speech recognizers that are part of freely accessible spoken information systems (see e.g. [6]). Since the way people pronounce foreign words depends to a large extent on their mother tongue, it seems reasonable to assume that non-native speech recognition can be improved by means of language adaptation. A related issue is the recognition of foreign words, especially foreign names. This issue has received increasing attention recently (see e.g. [57, 73]) because many current applications contain foreign names, and the recognition of these foreign names appears to be problematic. Also in this case adaptation can lower the error rates.

Another example, one that is maybe less known in the ASR community, is ASR for people with a speech handicap, e.g. ASR for people with dysarthria [22, 52, 58, 59, 89]. Since the pronunciations of individuals with speech disabilities often differ substantially from the canonical pronunciations present in a standard ASR lexicon, lexicon adaptation could be beneficial in many cases. It is interesting to note that speaker adaptation for individuals with a speech handicap is often tackled from two sides: (1) the speech recognizer is adapted to the speaker, and (2) the speaker is adapted to the speech recognizer through speech training [52, 59]. Although it is known that many speakers change the way they articulate when addressing a speech recognizer, adaptation of the speaker to the speech recognizer is generally not viewed as a viable method to do speaker adaptation. The reason why this approach is used for individuals with a speech handicap is probably that in these cases standard dictation systems are used, which allow adaptation of the acoustic models by means of speaker enrollment and in certain cases even addition of entries to the lexicon, but they do not allow changing the pronunciations present in the lexicon. The pronunciations in the lexicon often differ substantially from the pronunciations of these individuals, which results in a decrease in performance. Since the pronunciations in the lexicon cannot be adapted, the individuals are trained to change their pronunciations in order to increase recognition performance. Although, at a first glance, this may not seem a user friendly approach, the advantage it offers is that the intelligibility of these individuals can improve, which has already been observed.

Finally, the target can be constant or changing over time. To continue with the above mentioned example: some forms of dysarthria are progressive, and, consequently, the pronunciations of these individuals are not constant. Similarly, the pronunciations of users of CALL systems also change over time. At least, this is what one hopes. After all, the goal of such CALL systems is that learners acquire the foreign language and its correct pronunciation, and thus, hopefully, their pronunciation should gradually improve. In cases such as these, in which the target is not constant, the overall performance of the speech recognizer can probably be improved by repeating the adaptation regularly.

All the examples above illustrate situations in which lexicon adaptation is likely to be beneficial. This raises the question as to when lexicon adaptation should be applied. In most publications on lexicon adaptation this question is not addressed. It is obvious that the answer to this question depends on the differences between the pronunciations of the speech signals (utterances of an individual or a group of individuals) and the pronunciations contained in the lexicon. The following three properties of these differences seem to be important: (1) frequency, (2) magnitude, and (3) description of these differences. After all, it is more likely that lexicon adaptation can improve the performance of a speech recognizer if the following three criteria are met: (1) the differences occur frequently, (2) the differences are large, and (3) the differences can be expressed in terms of the acoustic-phonetic symbols present in the lexicon.

Not all of the cases discussed in this section received equal attention in the past. For instance, there are more publications on automatic off-line adaptation in which the starting point is a lexicon with canonical transcriptions which is optimized for a certain corpus, than publications on the other cases of lexicon adaptation. It is inevitable then that the overview below contains

many references to studies about the first type of adaptation, and fewer references to studies about other types of adaptation.

3. Lexicon adaptation: Overview of approaches

In this section an overview is presented of the approaches to lexicon adaptation that have been proposed so far. First, in section 3.1, I discuss the types of pronunciation variation that have been modeled. Since all methods require the following two steps:

1. finding information on pronunciation (variation), and
2. using this information in ASR,

these two aspects are addressed separately in sections 3.2 and 3.3, respectively.

3.1. Type of pronunciation variation

An important distinction that is often drawn in pronunciation variation modeling is that between within-word variation and cross-word variation. Within-word variation is typically the sort of variation that can easily be modeled at the level of the lexicon by simply adding pronunciation variants [1, 2, 3, 4, 7, 8, 9, 10, 14, 16, 26, 27, 29, 31, 39, 42, 43, 44, 48, 49, 54, 56, 60, 61, 62, 72, 74, 75, 83, 86, 88, 91, 95, 97, 98]. Besides within-word variation, cross-word variation also occurs, especially in continuous speech. Therefore, cross-word variation should also be accounted for. A solution combining the ease of modeling at the level of the lexicon and the need to model cross-word variation is the use of multi-words [7, 27, 48, 64, 70, 72, 86]. In this approach, sequences of words (usually called multi-words) are treated as one entity in the lexicon and the variations that result when the words are strung together are modeled by including different variants of the multi-words in the lexicon. It is important to note that, in general, with this approach only a small portion of cross-word variation is modeled, e.g. the variation occurring between words that occur in very frequent sequences. Besides the multi-word approach, other methods have been proposed to model cross-word variation such as those described in [3, 8, 9, 14, 16, 60, 67, 70, 79, 83, 96, 97].

Given that both within-word variation and cross-word variation occur in running speech, it is necessary to model both types of variation. This has already been done in [7, 8, 9, 14, 16, 27, 48, 60, 74, 83, 86, 97].

3.2. How to obtain the information on pronunciation

An important step in lexicon adaptation is finding information on pronunciation variation. Although much of this information can be found in the literature, this generally appears turns out to be insufficient for various reasons, the most important being that it mostly concerns laboratory speech instead of spontaneous speech. In the sections below, I briefly examine two different approaches for gathering information on pronunciation variation: knowledge-based and data-driven methods.

3.2.1. Knowledge-based

In knowledge-based studies, information on pronunciation variation is primarily derived from sources that are already available [1, 3, 10, 14, 21, 26, 27, 48, 50, 51, 54, 56, 60, 64, 67, 70, 76, 79, 83, 92, 96, 98]. The existing sources can be pronunciation dictionaries (see e.g. [76]), or rules on pronunciation variation from linguistic studies. In general, these rules are optional phonological rules concerning deletions, insertions and substitutions of phones [1, 3, 12, 14, 26, 27, 48, 54, 56, 60, 64, 67, 70, 79, 83, 96, 98]. A possible drawback of knowledge-based methods is that these sources usually only provide qualitative information about the pronunciations, and no quantitative information. Since quantitative information is needed for ASR, it has to be obtained from the acoustic signals. Furthermore, not many suitable pronunciation dictionaries do exist. Finally, as mentioned above, most of the available sources contain information on the variations that occur in laboratory speech, whereas information concerning spontaneous speech is generally lacking.

3.2.2. Data-driven

The idea behind data-driven methods is that information on pronunciation variation is obtained directly from the speech signals [2, 4, 8, 9, 16, 29, 31, 37, 38, 39, 41, 42, 43, 44, 49, 61, 62, 64, 69, 72, 74, 75, 86, 88, 91, 95, 97]. To this end, the acoustic signals are analyzed in order to determine the different ways in which the words are realized. A common stage in this analysis is transcribing the acoustic signals. Transcriptions of the acoustic signals can be obtained either manually [21, 29, 37, 38, 39, 61, 74, 75, 96] or (semi-) automatically [1, 2, 4, 7, 16, 31, 42, 48, 49, 56, 62, 72, 74, 83, 88, 91, 95, 97]. The latter is usually done either with a phone(me) recognizer [2, 31, 62, 72, 91, 95] or by means of forced recognition (which is also referred to as forced alignment or Viterbi alignment) [1, 4, 7, 16, 48, 49, 56, 74, 83, 97].

The transcriptions can simply be stored in a list, and a selection of them can be added to the lexicon (see section 3.3). However, usually formalizations are derived from the data-driven transcriptions [2, 16, 20, 31, 44, 49, 72, 91, 97]. In general, this is done in the following manner. The transcriptions of the utterances are aligned with the corresponding canonical transcription (obtained by concatenating the canonical transcriptions of the individual words). Alignment is done by means of a Dynamic Programming (DP) algorithm [16, 29, 31, 39, 49, 72, 74, 91, 95, 96, 97]. The resulting DP-alignments can then be used to:

- derive rewrite rules [2, 16, 49, 72, 97],
- train an artificial neural network [20, 31],
- train decision trees [29, 74], and to
- calculate a phone confusion matrix [91].

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

3.3. How to use the information on pronunciation

In the previous section an overview was given of the various ways in which information on pronunciation can be obtained. I here proceed to describe how the information thus obtained can be used for ASR. Pronunciation adaptation at the level of the lexicon is usually done by adding pronunciation variants, and their transcriptions, to the lexicon [1, 3, 7, 10, 14, 16, 21, 26, 27, 31, 42, 43, 44, 48, 49, 54, 56, 60, 62, 64, 72, 74, 76, 86, 91, 95, 96, 97, 98]. In order to be able to add pronunciation variants to the lexicon, these pronunciation variants first have to be generated. A discussion of this stage is offered in the following section.

3.3.1. Variant generation

The pronunciation variants can be generated manually [3, 74] or selected from specific lists [28]. However, they usually are generated automatically by means of various procedures:

- rules [1, 2, 3, 14, 16, 28, 48, 49, 60, 64, 72, 83, 97],
- artificial neural networks [31],
- grapheme-to-phoneme converters [56],
- phone(me) recognizers [62, 64, 72, 86, 95],
- optimization with maximum likelihood criterion [42, 43], and
- decision trees [29, 74].

3.3.2. Variant selection

The rationale behind adding pronunciation variants to the lexicon is that, with multiple transcriptions of the same word, the chance is increased that the speech recognizer selects a transcription belonging to the correct word (for an incoming, unknown signal). In turn, this should lead to lower error rates. However, adding pronunciation variants to the lexicon usually also introduces new errors because the 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 [16, 31, 42, 43, 44, 48, 49, 56, 62, 64, 74, 86, 91, 97]. For this purpose, different criteria can be used, such as:

- frequency of occurrence of the variants [48, 49, 72, 74, 83, 95],
- a maximum likelihood criterion [42, 43, 44],
- confidence measures [86], and
- the degree of confusability between the variants [86, 91].

A description of a method to detect confusable pairs of words or transcriptions is given in [77]. If rules are used to generate pronunciation variants, then certain rules can be selected (and others discarded), as in [16, 56, 83, 97] where rules are selected on the basis of their frequency and application likelihood.

4. Discussion

Automatic speech recognizers make errors, and the number of errors is especially large when conversational, extemporaneous speech has to be recognized. These errors have various sources: pronunciation variation, noise, training-recognition mismatch, etc. Therefore, there are also various kinds of adaptation techniques that try to resolve (part of) these errors. In general, it is not known what the contribution of each of these different error sources is, and thus it is impossible to known beforehand how much improvement can be obtained with each of these adaptation techniques. Furthermore, the contribution of each error source will differ across situations and interaction effects between the sources are also possible. For instance, it is well-known that if there is a lot of background noise, people will often change the way they pronounce speech sounds (Lombard effect). In such a case, both the background noise itself and the resulting pronunciation variation can cause recognition errors (see e.g. [46, 47]). Finally, there can be an overlap in the errors that are resolved by the different adaptation techniques.

In this paper, we have looked at the errors that can be resolved by adapting the lexicon. Usually this is done by adding pronunciation variants to the lexicon. However, previous results have shown that only adding pronunciation variants to the lexicon during recognition is sub-optimal. Better results are generally obtained when also the probabilities of the pronunciation variants are taken into account (either in the lexicon or in the language model), and sometimes retraining the acoustic models results in an extra improvement [48, 49, 54]. In short, the best results are generally found when pronunciation variants are used during training and recognition, at all levels of the speech recognizer: lexicon, acoustic models and language models. Furthermore, a problem with this method is that certain words have numerous variants with very different frequencies of occurrence. Some quantitative data on this phenomenon can be found in [37]. For instance, if we look at the data for the word 'that', (in Table 2 on page 50 of [37]) we can see that this word appears 328 times in the corpus used by Greenberg, that it has 117 different pronunciations and that the most frequent variant only covers 11% of all pronunciations. 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, in many cases the frequency of the pronunciation variants gradually diminishes, and thus there is no clear distinction between frequent and infrequent variants. Furthermore, a variant can be frequent because it occurs often (absolute frequency), or because it occurs frequently given the number of times the word occurs (relative frequency). Besides absolute and relative frequency, other selection criteria have been used, like e.g. entropy and likelihood.

Whatever selection procedure is used, the pronunciation variants for the lexicon first have to be generated. A lexicon contains phonetic units, and the pronunciations in the lexicon are transcriptions in terms of these units (a computer phonetic alphabet). One could wonder what the optimal transcription of a word or an utterance is. Previous studies have shown that the true human transcription does not exist (see e.g. [17]). Human labelers disagree on the transcriptions in a large number of cases. For instances, in transcriptions of Switchboard they disagree on the identity of the surface forms in more than 20% of the cases [80]. In [93, 94] an experiment is described in which 9 labelers were asked to decide for 467 cases whether a segment (a phone) was present or not. All 9 labelers agreed on only 246 (53%) of the cases [93, 94].

In the overview presented in section 3, it was pointed out that a speech recognizer is often used to make automatic transcriptions. These automatic transcriptions are usually evaluated by comparing them to human transcriptions. This procedure is usually applied because there is no better alternative, but it is clear that it is questionable for at least two reasons. First, since there is no human transcription that can be considered correct it is unclear with which human transcription a machine transcription should be compared. Second, whether a human tran-

scription is the optimal transcription for a lexicon of a speech recognizer is a moot point. After all, human speech recognition is substantially different from ASR. In order to understand what has been said, humans use many knowledge sources that the speech recognizer does not have at its disposal. And even then humans make recognition errors, which they often will try to correct by means of verbal and non-verbal communication. Furthermore, using transcriptions that were obtained by the ASR system itself, instead of human transcriptions, has the advantage that the transcriptions are more in line with the phone strings obtained later during decoding with the same ASR system [74]. To summarize, there is no straightforward way of determining what the optimal transcriptions for the lexicon are and how they should be obtained.

An important reason why making a transcription is so problematic, is that it requires chopping up the speech signal in consecutive parts that have to be labeled with (phonetic) symbols, i.e. a mapping should be made from a continuous acoustic space to discrete phonetic units. However, there are no clear boundaries in acoustic space. If we look at many pronunciations of the same word (or part of a word), we can see that changes are gradual, and not of a quantal nature. Already in 1956, Peterson and Barney showed that there is a large overlap between formant values of different vowels [68]. One might argue that formant values are not the optimal features for ASR, and that the picture might be different when features are used that are more common in ASR nowadays. However, SaraHlar recently showed that when Perceptual Linear Predictive (PLP, see [40]) coefficients are used, the changes in acoustic (spectral) space are also of a gradual nature [80: pp. 44-46, 81]. He looked at different realizations of a baseform phoneme, which he called /b/, which was realized as a surface form phone, named [s]. His results show that the acoustics of these different realizations are scattered around the average realization of the phone [s], with a bias towards the average realization of the phoneme /b/. In other words, a baseform phoneme /b/ does not suddenly change to a surface form [s], this is a gradual, partial process. Consequently, neither models for /b/ or [s] will provide a good fit for these realizations.

Changes are gradual, not only between different realizations of a word, but also within a single realization. Articulators cannot suddenly jump from one position to another, and thus the articulated speech sound changes gradually. Keeping this 'gradual nature of pronunciations' in mind, let us look at the assumptions made in 'standard ASR systems'. 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 [35: pp.16-17]) has proved to be untenable. In non-linear, autosegmental phonology [35, 36] 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, Deng and his colleagues have built ASR systems with which promising results have been obtained [19].

Other important assumptions of 'standard ASR systems' are that consecutive frames are independent, which obviously is not the case, and that the feature values can be calculated locally. Although some dynamic information can be obtained from the derivatives of these features, the problem remains that the analysis window on which feature values are calculated is very small, whereas it is known from research on human perception that for perceiving one speech sound subjects rely on information contained in adjacent sounds.

Given this discrepancy between the data (and its properties) on the one hand, and current speech recognizers and the underlying assumptions on the other hand, one might wonder whether the problem of lexicon adaptation is an ill-defined problem, and whether adaptation of current speech recognizers is the right way to proceed. To come back to the quote by Darwin at the beginning of this paper, we have witnessed a gradual variability in a favorable direction in the sense that speech recognizers have gradually become better. However, we still have a long way to go. In his masterpiece on evolution Darwin assumes that the adaptation process is a gradual process. Nowadays, most scientists believe that periods of gradual changes (of an evolutionary nature) alternate with periods of sudden, drastic changes (of a revolutionary nature). Maybe it is time for a revolution in ASR land.

5. Acknowledgments

I would like to thank Catia Cucchiarini for useful discussions and constructive comments on previous versions of this article.

6. References

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

[2] Amdal, I., Korkmazskiy, F., and Surendran, A.C., 2000. Joint pronunciation modelling of non-native speakers using data-driven methods. In: Proc. of ICSLP-2000, Beijing, China, pp. 622-625.

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

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

[5] Barnett, J., 1974. A phonological rule compiler. In: Proc. of 'the IEEE Symposium on Speech Recognition' (see [23]), pp. 188-192.

[6] Bartkova, K., Jouvet, D., 1999. Language based phone model combination for ASR adaptation to foreign accent. In: Proc. ICPhS-99, San Francisco, USA, August 1-7, 1999, vol.3, pp. 1725-1728.

[7] 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: Proc. of the ESCA Workshop 'Modeling Pronunciation Variation for Automatic Speech Recognition' (see [87]), pp. 13-16.

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

[9] Blackburn, C.S., Young, S.J., 1996. Pseudo-articulatory speech synthesis for recognition using automatic fea-

ture extraction from X-ray data. In: Proc. of ICSLP-96, Philadelphia, pp. 969-972.

[10] Bonaventura, P., Gallocchio, F., Mari, J., Micca, G., 1998. Speech recognition methods for non-native pronunciation variations. In: Proc. of the ESCA Workshop 'Modeling Pronunciation Variation for Automatic Speech Recognition' (see [87]), pp. 17-22.

[11] Carlson, R., Dunger, C., Granstrom, B., Oster, A. (Eds.), 1998. Proceedings of the ESCA Workshop 'STiLL: Speech Technology in Language Learning', Marholmen, Sweden, May 1998.

[12] Cohen, M., 1989. Phonological structures for speech recognition. Ph.D. thesis, Univ. of California, Berkeley, USA.

[13] Cohen, P.S., Mercer, R.L., 1974. The Phonological Component of an Automatic Speech Recognition System. In: Proc. of 'the IEEE Symposium on Speech Recognition' (see [23]), pp. 177-187.

[14] 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, Inc., New York, 1975, pp. 275-320.

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

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

[17] Cucchiarini, C., 1993. Phonetic transcription: a methodological and empirical study. Ph.D. thesis, Univ. of Nijmegen, The Netherlands.

[18] Darwin, C., 1958. The Origin of Species by Means of Natural Selection, Penguins Books, New York, Edition 6, printed 1958.

[19] Deng, L., Sun, D., 1994. A statistical approach to ASR using the atomic speech units constructed from overlapping articulatory features. J. Acoust. Soc. Amer., 95 (5), May 1994, 2702-2719.

[20] Deshmukh, N., Weber, M., Picone, J., 1996. Automated generation of N-best pronunciations of proper nouns. In: Proc. of ICASSP-96, Atlanta, pp. 283-286.

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

[22] Doyle, P.C., Leeper, H.A., Kotler, A., Thomas-Stonell, N., O'Neill, C., Dylke, M., Rolls, K., 1997. Dysarthric speech: A comparison of computerized speech recognition and listener intelligibility. Journal of Rehabilitation Research and Development, 34 (3), pp. 309-316.

[23] Erman, L., 1974. Proc. of the IEEE Symposium on Speech Recognition, Carnegie Mellon Univ., Pittsburgh Pa., 15-19 April 1974, 295 pages. (IEEE Catalog No. 74CH0878-9 AE).

[24] Eskenazi, M., 1993. Trends in speaking styles research. In: Proc. of Eurospeech-93, Berlin, pp. 501-509.

[25] Eskenazi, M., 1999. Using automatic speech processing for foreign language pronunciation tutoring: Some issues and a prototype. Language Learning & Technology , Vol. 2, No. 2, January 1999, pp. 62-76.

[26] Ferreiros, J., Macas-Guarasa, J., Pardo, J.M., Villarrubia, L., 1998. Introducing multiple pronunciations in Spanish speech recognition systems. In: Proc. of the ESCA Workshop 'Modeling Pronunciation Variation for Automatic Speech Recognition' (see [87]), pp. 29-34.

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

[28] Flach, G., 1995. Modelling pronunciation variability for spectral domains. In: Proc. of Eurospeech-95, Madrid, pp. 1743-1746.

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

[30] Friedman, J., 1974. Computer exploration of fast speech rules. In: Proc. of 'the IEEE Symposium on Speech Recognition' (see [23]), pp. 197-203.

[31] Fukada Toshiaki, Yoshimura Takayoshi, Sagisaka Yoshinori, 1999. Automatic generation of multiple pronunciations based on neural networks. Speech Communication, 27 (1), pp. 63-73.

[32] Giles, H., Powesland, P., 1975. Speech style and social evaluation. Cambridge University Press, Cambridge.

[33] Giles, H., Smith, P., 1979. Accommodation theory: Optimal levels of convergence. In: Giles, H., stClair, R. (Eds.) Language and social psychology, Blackwell, Oxford.

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

[35] Goldsmith, J., 1976. Autosegmental phonology. Doctoral thesis, Massachussets Institute of Technology, Cambridge. [Bloomington, Indiana: Indiana University Linguistics Club. New York: Garland Press, 1979].

[36] Goldsmith, J.A., 1990. Autosegmental and Metrical Phonology. Oxford: Blackwell.

[37] Greenberg, S., 1998. Speaking in shorthand - A syllable-centric perspective for understanding pronunciation variation. In: Proc. of the ESCA Workshop 'Modeling Pronunciation Variation for Automatic Speech Recognition' (see [87]), pp. 47-56.

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

[39] Heine, H., Evermann, G., Jost, U., 1998. An HMM-based probabilistic lexicon. In: Proc. of the ESCA Workshop 'Modeling Pronunciation Variation for Automatic Speech Recognition' (see [87]), pp. 57-62.

[40] Hermansky, H., 1990. Perceptual linear predictive (PLP) analysis of speech. J. Acoust. Soc. Amer., 87 (4), pp. 1738-1752.

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

[42] Holter, T., 1997. Maximum Likelihood Modelling of Pronunciation in Automatic Speech Recognition. Ph.D. thesis, Norwegian University of Science and Technology.

[43] Holter, T., Svendsen, T., 1999. Maximum likelihood modelling of pronunciation variation. Speech Communication, 29 (2-4), pp. 177-191.

[44] Imai, T., Ando, A., Miyasaka, E., 1995. A New Method for Automatic Generation of Speaker-Dependent Phonological Rules. In: Proc. of ICASSP-95, Detroit, pp. 864-867.

[45] Jelinek, F., Bahl, L.R., Mercer, R.L., 1974. Design of a linguistic statistical decoder for the recognition of continuous speech. In: Proc. of 'the IEEE Symposium on Speech Recognition' (see [23]), pp. 255-260.

[46] Junqua, J-C., 1993. The Lombard Reflex and its role on human listeners and automatic speech recognisers. J. Acoust. Soc. Amer., 93 (1), pp. 510-524.

[47] Junqua, J-C., 1999. The Lombard effect: A reflex to better communicate with others in noise. In: Proc. of ICASSP-99, Phoenix, pp. 2083-2086.

[48] 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), pp. 193-207.

[49] Kessens, J.M., Strik, H., Cucchiarini, C. 2000. A bottom-up method for obtaining information about pronunciation variation. In: Proc. of ICSLP-2000, Beijing, China.

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

[51] Kipp, A., Wesenick, M.-B., Schiel, F., 1997. Pronunciation Modeling Applied to Automatic Segmentation of Spontaneous Speech. In: Proc. of EuroSpeech-97, Rhodes, pp. 1023-1026.

[52] Kotler, A., Thomas-Stonell, N., 1997. Effects of speech training on the accuracy of speech recognition for an individual with a speech impairment. Augmentative and Alternative Communication, 13, pp. 71-80.

[53] Labov, W., 1972. Sociolinguistic patterns. University of Pennsylvania Press, Philadelphia.

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

[55] Laver, J., 1994. Principles of Phonetics. Cambridge University Press, Cambridge.

[56] Lehtinen, G., Safra, S., 1998. Generation and selection of pronunciation variants for a flexible word recognizer. In: Proc. of the ESCA Workshop 'Modeling Pronunciation Variation for Automatic Speech Recognition' (see [87]), pp. 67-72.

[57] Lindström, A., Kasaty, A., 2000. A two-level approach to the handling of foreign items in Swedish speech technology applications. In: Proc. of ICSLP-2000, Beijing, China.

[58] Magnuson, T., and Blomberg, M., 2000. Acoustic analysis of dysarthric speech and some implications for ASR. KTH TMH-QPSR 1/2000, pp. 19-30.

[59] Manasse, N. J., Hux, K., Rankin-Erickson, J. L., 2000. Speech recognition training for enhancing the written language generation of a traumatic brain injury survivor. Brain Injury, 14, pp. 1015-1034.

[60] Mercer, R., Cohen, P., 1987. A method for efficient storage and rapid application of context-sensitive phonological rules for ASR. IBM J. Res. Develop., Vol. 31, No. 1, January 1987, pp. 81-90.

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

[62] Mokbel, H., Jouvet, D., 1998. Derivation of the optimal phonetic transcription set for a word from its acoustic realisations. In: Proc. of the ESCA Workshop 'Modeling Pronunciation Variation for Automatic Speech Recognition' (see [87]), pp. 73-78.

[63] Murray, I.R., Arnott, J.L., 1993. Toward the simulation of emotion in synthetic speech: A review of the literature on human vocal emotion. J. Acoust. Soc. Amer., 93 (2), pp. 1097-1108.

[64] Nock, H.J., Young, S.J., 1998. Detecting and correcting poor pronunciations for multiword units. In: Proc. of the ESCA Workshop 'Modeling Pronunciation Variation for Automatic Speech Recognition' (see [87]), pp. 85-90.

[65] O'Malley, M.H., Cole, A., 1974. Testing phonological rules. In: Proc. of 'the IEEE Symposium on Speech Recognition' (see [23]), pp. 193-196.

[66] Oshika, B.T., Zue, V.W., Weeks, R.V., Neu, H., 1974. The role of phonological rules in speech understanding research. In: Proc. of 'the IEEE Symposium on Speech Recognition' (see [23]), pp. 204-207.

[67] Perennou, G., Brieussel-Pousse, L., 1998. Phonological component in ASR. In: Proc. of the ESCA Workshop 'Modeling Pronunciation Variation for Automatic Speech Recognition' (see [87]), pp. 91–96.

[68] Peterson, G.E. and Barney, H.L., 1952. Control methods used in a study of the vowels. J. Acoust. Soc. Amer., 24 (2), pp. 175-184.

[69] Polzin, T.S., Waibel, A.H., 1998. Pronunciation variations in emotional speech. In: Proc. of the ESCA Workshop 'Modeling Pronunciation Variation for Automatic Speech Recognition' (see [87]), pp. 103-108.

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

[71] Rabinowitz, A.S., 1974. Phonetic to graphemic transformation by use of a stack procedure. In: Proc. of 'the IEEE Symposium on Speech Recognition' (see [23]), pp. 212-217.

[72] Ravishankar, M., Eskenazi, M., 1997. Automatic generation of context-dependent pronunciations. In: Proc. of EuroSpeech-97, Rhodes, pp. 2467-2470.

[73] Riis, S.K., Pedersen, M.W., Jensen, K.J., 2001. Multilingual text-to-phoneme mapping. In: Proc. of Eurospeech-2001, Aalborg, Danmark.

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

[75] Ristad, E.S., Yianilos, P.N., 1998. A surficial pronunciation model. In: Proc. of the ESCA Workshop 'Modeling Pronunciation Variation for Automatic Speech Recognition' (see [87]), pp. 117-120.

[76] Roach, P., Arnfield, S., 1998. Variation information in pronunciation dictionaries. In: Proc. of the ESCA Workshop 'Modeling Pronunciation Variation for Automatic Speech Recognition' (see [87]), pp. 121-124.

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

[78] Rovner, P., Makhoul, J., Wolf, J., Colarusso, J., 1974. Where the words are: lexical retrieval in a speech understanding system. In: Proc. of 'the IEEE Symposium on Speech Recognition' (see [23]), pp. 160-164.

[79] Safra, S., Lehtinen, G., Huber, K., 1998. Modeling pronunciation variations and coarticulation with finite-state transducers in CSR. In: Proc. of the ESCA Workshop 'Modeling Pronunciation Variation for Automatic Speech Recognition' (see [87]), pp. 125-130.

[80] SaraHlar, M., 2000. Pronunciation Modeling for Conversational Speech Recognition. Ph.D. thesis, Johns Hopkins University, Baltimore, MD, USA.

[81] SaraHlar, M. and Khudanpur, S., 2000. Pronunciation Ambiguity vs. Pronunciation Variability in Speech Recognition. ICASSP-2000, Istanbul, Turkey.

[82] Scherer, K.R., Giles, H., 1979. Social Markers in Speech. Cambridge: Cambridge University Press.

[83] Schiel, F., Kipp, A., Tillmann, H.G., 1998. Statistical modelling of pronunciation: it's not the model, it's the data. In: Proc. of the ESCA Workshop 'Modeling Pronunciation Variation for Automatic Speech Recognition' (see [87]), pp. 131-136.

[84] Shockey, L., Erman, L.D., 1974. Sub-lexical levels in the HEARSAY II speech understanding system. In: Proc. of 'the IEEE Symposium on Speech Recognition' (see [23]), pp. 208-210.

[85] Singh, R., Raj, B., and Stern, R. M., 2000. Structured Redefinition of Sound Units by Merging and Splitting for Improved Speech Recognition. In: Proc. of ICSLP-2000, Beijing, China.

[86] Sloboda, T., Waibel, A., 1996. Dictionary Learning for Spontaneous Speech Recognition. In: Proc. of ICSLP-96, Philadelphia, pp. 2328-2331.

[87] Strik, H., Kessens, J.M., Wester, M. (Eds.), 1998. Proceedings of the ESCA Workshop 'Modeling Pronunciation Variation for Automatic Speech Recognition', Rolduc, Kerkrade, The Netherlands, 4-6 May 1998. A2RT, University of Nijmegen, 168 pages.

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

[89] Talbot, N., 2000. Improving the speech recognition in the ENABL project. KTH TMH-QPSR 1/2000, pp. 31-38.

[90] Tappert, C. C., 1974. Experiments with a tree search method for converting noisy phonetic representation into standard orthography. In: Proc. of 'the IEEE Symposium on Speech Recognition' (see [23]), pp. 261-266.

[91] Torre, D., Villarrubia, L., Hernndez, L., Elvira, J.M., 1997. Automatic Alternative Transcription Generation and Vocabulary Selection for Flexible Word Recognizers. In: Proc. of ICASSP-97, Munich, pp. 1463-1466.

[92] Wesenick, M.-B., 1996. Automatic generation of German pronunciation variants. In: Proc. of ICSLP-96, Philadelphia, pp. 125-128.

[93] Wester, M., Kessens, J.M., 1999. Comparison Between Expert Listeners and Continuous Speech Recognizers in Selecting Pronunciation Variants. In: Proc. of ICPhS-99, San Francisco, USA, pp. 723-726.

[94] Wester, M., Kessens, J.M., Cucchiarini, C., Strik, H., 2001. Obtaining phonetic transcriptions: A comparison between expert listeners and a continuous speech recognizer. To appear in Language & Speech 44(4).

[95] Williams, G., Renals, S., 1998. Confidence measures for evaluating pronunciation models. In: Proc. of the ESCA Workshop 'Modeling Pronunciation Variation for Automatic Speech Recognition' (see [87]), pp. 151-156.

[96] Wiseman, R., Downey, S., 1998. Dynamic and static improvements to lexical baseforms. In: Proc. of the ESCA Workshop 'Modeling Pronunciation Variation for Automatic Speech Recognition' (see [87]), pp. 157-162.

[97] Yang, Q., Martens, J-P., 2000. Data-driven lexical modeling of pronunciation variations for ASR. In: Proc. ICSLP-2000, Beijing, pp. 417-420.

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