MODELING PRONUNCIATION VARIATION FOR ASR: OVERVIEW AND COMPARISON OF METHODS

Helmer Strik, Catia Cucchiarini

A²RT, Dept. of Language and Speech, University of Nijmegen
{strik, catia}@let.kun.nl, http://lands.let.kun.nl/TSpublic/{strik, catia}

ABSTRACT

In this contribution an overview is provided of the papers presented at this workshop. 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.

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. This variation in pronunciation is a major problem in 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. In general, the consequence is that they also 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. On the contrary, pausing between words is highly unnatural. Therefore, attempts were made in ASR research to improve technology so that it could handle less artificial speech. As a consequence, the type of speech used in ASR research has gradually progressed from isolated words to connected words, carefully read speech, and finally 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. As a matter of fact, there has been an increase in the amount of research on this topic (see e.g. [40]), which is evident from the growing number of contributions to conferences, and also from the organization of this workshop.

The aim we had in mind when we decided to organize this meeting was to create the opportunity for researchers working on this topic to have in-depth discussions on the problem of pronunciation variation and its possible solutions. Moreover, we thought it would be very interesting if, on the basis of these discussions, it were possible to draw some conclusions as to the best way in which to approach the pronunciation variation modeling problem. This would require an objective comparison of the methods proposed by the various authors.

To pave the way for this kind of discussion at the workshop, this paper provides an overview of the methods that will be presented at this meeting. The presentation of the various methods will be organized around some of the major characteristics that distinguish pronunciation variation modeling techniques from each other. In illustrating these characteristics we will not limit ourselves to the contributions to this workshop, but, where necessary, reference will be made to related research that has been presented previously.

After having presented the different techniques, we will address the issues of evaluation and comparison, 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

In choosing which method to use for pronunciation variation modeling a number of decisions have to be made. 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 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. Below we will consider these questions and the possible answers in more detail.

2.1. Type of pronunciation variation

With respect to the type of pronunciation variation to be modeled the choice is between variation within words and variation across word boundaries. In general, this choice will be influenced by several factors such as the type of ASR and the language which is used, and the level at which modeling will take place.

Modeling within-word variation is an obvious choice if the ASR makes use of a lexicon with word entries, because in this case variants can simply be added to the lexicon. Given that almost all ASRs use such a lexicon, within-word variation is modeled in the majority of the methods [1, 2, 3, 4, 5, 6, 7, 8, 10, 11, 12, 13, 17, 20, 22, 23, 24, 25, 33, 36]. However, there are methods that model cross-word variation such as [5, 16, 22, 26, 27, 28].

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 [3, 15, 19, 24, 39]. In this approach sequences of words (usually called multi-words) are treated as one entity in the lexicon (see section 2.4.1.).

2.2. Information sources

Once decisions have been made as to the type of variation, it is important to choose the source from which information on pronunciation variation will be retrieved. In this regard a distinction can be drawn between data-driven vs. knowledge-based methods.

In data-driven methods the information on pronunciation variation is mainly obtained from the acoustic signals [2, 5, 7, 8, 9, 10, 11, 13, 17, 18, 19, 20, 25, 27, 28, 33]. In this type of approach the acoustic signals are usually first transcribed. Subsequently the transcriptions can be used for different purposes, as will be explained in section 2.3.

Transcriptions of the acoustic signals can be obtained either manually [7, 9, 10, 19, 20, 26, 27, 31] or automatically [2, 5, 8, 10, 12, 13, 15, 19, 25, 28, 33]. Given that acquiring manual transcriptions for very large corpora is extremely time-consuming, and therefore costly, the use of automatically obtained transcriptions is becoming more common. Moreover, there is another reason why transcriptions obtained automatically with the ASR itself could be beneficial, viz. that these transcriptions are more in line with the phone strings obtained later during recognition with the same ASR. This is also mentioned by Riley et al. [19] who conclude: "Further, our results indicate that while a handlabeled corpus is very useful as a bootstrapping device, estimates of pronunciation probabilities, context effects, *etc.*, are best derived from larger amounts of automatic transcriptions, preferably done using the same set of acoustic models which will eventually be used for recognition."

In knowledge-based studies information on pronunciation variation is primarily derived from sources that are already available [1, 4, 6, 12, 14, 16, 21, 22, 23, 24, 26, 31, 36, 37, 38, 39, 42]. In general, the way in which the information on pronunciation variation is represented varies for the different sources. It can be a formalized representation in terms of rules, as in linguistic studies, or enumerated information in terms of pronunciation forms, as in pronunciation dictionaries. These two types of representations are discussed in detail in section 2.3.

The distinction between the data-driven and the knowledge-based approaches is related to the difference 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, 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.

In this sense most studies presented at this workshop can be said to be either data-driven or knowledge-based, because for each of them it is possible to say what the starting point of the research was (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 the direction of the developing process is solely upward or downward, the flow of information can be 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, because, for example, 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 in ASR research the emphasis is now on spontaneous speech. However, the knowledge on pronunciation variation that can be found in the literature usually concerns other speech 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. To overcome this problem one can resort to a combination of top-down and bottom-up approaches, as explained above.

On the other hand, a possible disadvantage of data-driven studies is that for every new corpus the whole process of transcribing the speech material and deriving information on pronunciation variation has to be repeated. In other words, information obtained on the basis of data-driven studies does not generalize easily to situations other than the one in question.

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 [5, 8, 27, 28, 30, 33, 41]. 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 [5, 7, 8, 10, 26, 27, 28, 33, 41]. The resulting DP-alignments can then be used to

- derive rewrite rules [5, 27, 28]
- train an artificial neural network (ANN) [8, 30, 33]
- calculate a phone confusion matrix [41].

In these three cases the information about pronunciation variation present in the DP-alignments is formalized in terms of rewrite rules, 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 [1, 6, 12, 15, 16, 22, 23, 24, 26, 39]. Rules (either obtained from data or from linguistic studies) and ANNs are then used to generate the various pronunciation 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 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 decide 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. However, the use of formalizations also has some disadvantages, like overgeneration and undergeneration, owing to incorrect specifications of the rules applied, or overcoverage and undercoverage. Both types of problems should not arise when using enumerated information.

2.4. Level of modeling

Given that most ASRs 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. For example, variation modeling can happen in the lexicon and in the language model simultaneously, as will be described below.

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 [1, 3, 4, 5, 6, 8, 11, 12, 13, 15, 19, 21, 24, 25, 26, 27, 28, 31, 36, 41]. The rationale behind adding pronunciation variants to the lexicon 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 transcriptions of 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 of the ASR [5, 8, 11, 12, 13, 15, 19, 24, 27, 28, 36, 41]. For this purpose different criteria can be used, such as frequency of occurrence of the variants [24, 36], degree of confusability between the variants [41] or a maximum likelihood criterion [11].

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 could also be included in the lexicon. By using multi-words Beulen et al. [3] and Wester et al. [24] achieve a substantial improvement. On the other hand, Nock and Young [15] conclude that "No clear evidence of multi-words being beneficial was found under any of the selection criteria".

Before variants can be selected, they have to be obtained, in the first place. In the previous section we saw that variants can either be obtained directly (from data or available sources), or be generated by means of rewrite rules or ANNs. The contributions to this workshop contain examples of all methods.

Since rule-based methods are probably the methods used most often, it is interesting to note that Nock &

Young [15] conclude that "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.).

In [32] the 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 [2, 3, 4, 9, 10, 15, 19, 23, 24, 29, 34, 35, 36]. Optimization can be attained in different ways.

2.4.2.1. Iterative transcribing

An obvious way of optimizing the acoustic models is by using a procedure which we will refer to as iterative transcribing. In this procedure pronunciation variants are used both during training and recognition [3, 19, 23, 24, 36]. The goal of this procedure is the alternate improvement of the transcriptions contained in the training corpus and of the acoustic models trained on this corpus. The transcriptions available and a canonical lexicon are the starting points. These are used to train the first set of acoustic models. Subsequently, the pronunciation variants are added to the lexicon. For every word in the corpus for which pronunciation variants are present in the lexicon, the ASR itself selects the optimal one. In this way new, updated transcriptions are obtained which, in turn, are used to train new acoustic models. Updating the transcriptions and re-training the acoustic models can be repeated iteratively.

In general, this procedure seems to improve the performance of the ASR [19, 23, 24, 36]. However, Beulen et al. [3] found that in some cases the performance does not improve, but remains unchanged or even deteriorates. Furthermore, the beneficial effect of including pronunciation variants during recognition is usually larger than that deriving from iterative transcribing. In spite of this, it seems worthwhile to test iterative transcribing 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.

2.4.2.2. Other basic units

In most ASRs 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 [29, 34]. In [29] 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 [34] is a set of symbols for (allo-)phones and sub-phonemic segments. 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 will decide 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, like e.g. (demi-)syllables [9, 10] 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 [9, 10], the syllable would seem to be suited as the basic unit of recognition. Greenberg [9] 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. 10]. 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 ASRs 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 [2, 35]. In both [2] and [35] the optimization is done with a maximum likelihood criterion.

In [9] no tests are described. For the syllable models in [10] the resulting levels of performance are lower than those of standard ASRs. Furthermore, in [2, 29, 34, 35] the observed levels of performance are comparable to those of phone-based ASRs (usually for limited tasks). Although these results are promising, it remains to be seen whether these methods are more suitable for modeling pronunciation variation than standard phone-based ASRs, 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) [5, 8, 12, 16, 23, 24, 27, 28, 30, 39]. 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 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 to add 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 first solution is to simply use the variants themselves (instead of the underlying words) to calculate the N-grams [23, 24]. For this procedure a transcribed corpus is needed which contains information about the realized pronunciation variants. Such a corpus can be obtained in a data-driven manner (see section 2.2.) or by the procedure of iterative transcribing (see section 2.4.2.1.). The goal of this method is to find the string of variants V which maximizes P(X|V)*P(V).

Method 2. A second solution would be 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 [5, 8, 16, 27, 28, 39]. The unigram determines the probability of a variant given the word, while the higher-order N-grams (i.e. N > 1) describe the probabilities of sequences of words. In this case the unigram 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 [5, 12]. Likewise, if an ANN is used to generate pronunciation variants, the ANN itself can produce probabilities of the pronunciation variants [8, 30].

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 first method than for the second. This could be a disadvantage of the first 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 second method the context-dependence of pronunciation variants cannot be modeled. This can be a disadvantage as pronunciation variation is often context-dependent, e.g. liaison in French [16, 39]. Within the second 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 [16, 39]. The probability of a pronunciation variant for a certain class is then represented in the unigram, while the probability of sequences of word classes is stored in the higher-order N-grams.

3. EVALUATION AND COMPARISON

In the previous section the various methods of modeling pronunciation variation have been described according to their major properties. In this presentation the emphasis was on the various characteristics of the methods, and not so much on their merits. This is not to say that the effectiveness of a method is not important. On the contrary, the extent to which each method achieves the goal it was intended for, be it reducing the number of errors caused by pronunciation variation or 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.

Although studies that provide insight into the processes underlying pronunciation variation are very useful (e.g. [9, 17]), the majority of the papers presented at this workshop focus on reducing word error rate (WER) by modeling pronunciation variation. The effectiveness of studies of this kind 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 proceedings) differ from almost none (and occasionally even a deterioration) to substantial ones.

In trying to draw general conclusions as to the effectiveness of the various methods one is then 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 gaining more insight into the way in which pronunciation variation can best be approached. 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, 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. On the contrary, they 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 ASRs. In studies on pronunciation variation modeling many different corpora are used. The choice of a given corpus implies at the same time 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.

Very often the task or application also dictates the type of speech that will have to be recognized. 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. [3] 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 and Brieussel-Pousse [16, 39] 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 other languages in which there is less or no liaison.

3.2. Differences between ASRs

As we all know, not all ASRs are similar. A method that works well for a certain ASR, can be less successful with another ASR. This will already be the case for ASRs with a similar architecture (i.e. a 'standard ASR' with the common phone-based HMMs), but it will certainly be true for ASRs with totally different architectures. For instance, Cremelie and Martens [5] obtain large improvements with a rule-based method for their segment-based ASR. However, this does not imply that the same rule-based method will be equally successful for another type of ASR.

Moreover, a method can be successful with a given ASR, not so much because it models pronunciation variation in the correct way, but because it corrects for the peculiarities of the ASR. To illustrate this point let us assume that a specific ASR 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 only because it corrects the errors made by the ASR. Although one could argue that the error made by the ASR (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, but do not necessarily generalize to other systems.

Let us state clearly that being able to correct for the peculiarities of an ASR 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. This is probably why in [19] it is argued that the ASR 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. After all, the peculiarities of all ASRs are not the same.

3.3. Differences in the baseline system

Another reason why it is difficult to compare methods is related to the baseline systems (the starting points) used. In order to illustrate this point, let us first recall briefly what a common method of evaluation is in this field of research. First, the performance is calculated for the baseline system, say 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 WER_{end}. The absolute improvement then is:

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

This is usually expressed in relative terms:

$$\%rel = (WER_{begin} - WER_{end})/WER_{begin}$$

The measure %rel yields higher numbers than the measure

%abs, but even higher numbers can be obtained by using

$$\%rel_2 = (WER_{begin} - WER_{end})/WER_{end}$$

The last equation is generally considered to be less correct. Furthermore, for most people %rel is more in agreement with their intuition than $\%rel_2$, i.e. most people would say that an improvement from 10% to 5% WER is an improvement of 50% and not an improvement of 100%.

Whatever equation is used, it is clear that the outcome of the equation depends on two numbers: WER_{begin} and WER_{end}. In most studies a lot of work is done in order to decrease WER_{end}, and this 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 that was available at the beginning of the research, or an ASR 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, like e.g. errors in the canonical lexicon. During the 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 WER_{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 speech style (e.g. read speech). A method is then tested with a corpus that contains speech of another accent and another speech 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 WER_{end}, while relatively little effort is put in WER_{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.

3.4. 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 ASRs.

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. We should discuss what kind of corpora and evaluation criteria could be used for this purpose. The current workshop provides a good opportunity for this discussion.

ACKNOWLEDGMENTS

The research of Dr. H. Strik has been made possible by a fellowship of the Royal Netherlands Academy of Arts and Sciences.

REFERENCES

Note: The first 26 references below are papers from the current proceedings.

[1] M. Adda-Decker (1998) Pronunciation variants across systems, languages and speaking style. This proc.

[2] M. Bacchiani, M. Ostendorf (1998) Joint acoustic unit design and lexicon generation. This proc.

[3] K. Beulen, S. Ortmanns, A. Eiden, S. Martin, L. Welling, J. Overmann, H. Ney (1998) Pronunciation modelling in the RWTH large vocabulary speech recognizer. This proc.

[4] P. Bonaventura, F. Gallocchio, J. Mari, G. Micca (1998) Speech recognition methods for non-native pronunciation variations. This proc.

[5] N. Cremelie, J.-P. Martens (1998) In search of pronunciation rules. This proc.

[6] J. Ferreiros, J. Macías-Guarasa, J.M. Pardo, L. Villarrubia (1998) Introducing multiple pronunciations in spanish speech recognition systems. This proc.

[7] E. Fosler-Lussier, N. Morgan (1998) Effects of

speaking rate and word frequency on conversational pronunciations. This proc.
[8] T. Fukada, T. Yoshimura, Y. Sagisaka (1998) Automatic generation of multiple pronunciations based on neural networks and language. This proc.
[9] S. Greenberg (1998) Speaking in shorthand: a syllable-centric perspective on understanding pronunciation. This proc.
[10] H. Heine, G. Evermann, U. Jost (1998) An HMM-based probabilistic lexicon. This proc.
[11] T. Holter, T. Svendsen (1998) Maximum likelihood modelling of pronunciation variation. This proc.
[12] G. Lehtinen, S. Safra (1998) Generation and selection of pronunciation variants for a flexible word recognizer. This proc.
[13] H. Mokbel, D. Jouvet (1998) Derivation of the optimal phonetic transcription set for a word from its acoustic realisations. This proc.
[14] F. Mouria-Beji (1998) Context and speed dependent phonemic models for continuous speech recognition. This proc.
[15] H.J. Nock, S.J. Young (1998) Detecting and correcting poor pronunciations for multiword units. This proc.
[16] G. Perennou, L. Brieussel-Pousse (1998) Phonological component in automatic speech recognition. This proc.
[17] S.D. Peters, P. Stubley (1998) Visualizing speech trajectories. This proc.
[18] T.S. Polzin, A.H. Waibel (1998) Pronunciation variations in emotional speech. This proc.
[19] M. Riley, W. Byrne, M. Finke, S. Khudanpur, A. Ljolje, (1998) Stochastic pronunciation modeling from hand-labelled phonetic corpora. This proc.
[20] E.S. Ristad, P.N. Yianilos (1998) A surficial pronunciation model. This proc.
[21] P. Roach, S. Arnfield (1998) Variation information in pronunciation dictionaries. This proc.
[22] S. Safra, G. Lehtinen, K. Huber (1998) Modeling pronunciation variations and coarticulation with finite-state . This proc.
[23] F. Schiel, A. Kipp, H.G. Tillmann (1998) Statistical modelling of pronunciation: it's not the model, it's the data. This proc.
[24] M. Wester, J.M. Kessens, H. Strik (1998) Improving the performance of a Dutch CSR by modelling pronunciation variation. This proc.
[25] G. Williams, S. Renals (1998) Confidence measures for evaluating pronunciation models. This proc.
[26] R. Wiseman, S. Downey (1998) Dynamic and static improvements to lexical baseforms. This proc.

---

[27] N. Cremelie & J.P. Martens (1995) On the use of pronunciation rules for improved word recognition. Proc. of Eurospeech-95, Madrid, Spain, Vol. III, pp. 1747-1750.
[28] N. Cremelie & J.P. Martens (1997) Automatic rule-based generation of word pronunciation networks. Proc. of EuroSpeech-97, pp. 2459-2462.
[29] L. Deng & D. Sun (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, V95(5), May 1994, pp.2702-2719.
[30] N. Deshmukh, M. Weber & J. Picone (1996) Automated generation of N-best pronunciations of proper nouns. Proc. of ICASSP-96, Atlanta, Vol. 1, pp. 283-286.
[31] S. Downey & R. Wiseman (1997) Dynamic and static improvements to lexical baseforms. Proc. of Eurospeech-97, Vol. 2, pp. 1027-1030.
[32] G. Flach (1995) Modelling pronunciation variability for spectral domains. Proc. of Eurospeech-95, Madrid, Vol. III, pp. 1743-1746.
[33] T. Fukada & Y. Sagisaka (1997) Automatic generation of a pronunciation dictionary based on a pronunciation network. Proc. of EuroSpeech-97, Rhodos, Vol. 5, pp. 2471-2474.
[34] J.J. Godfrey, A. Ganapathiraju, C.S. Ramalingam & J. Picone (1997) Microsegment-based connected digit recognition. Proc. of ICASSP-97, Munich, Vol. 3, pp. 1755-1758.
[35] T. Holter (1997) Maximum Likelihood Modelling of Pronunciation in Automatic Speech Recognition PhD thesis, Norwegian University of Science and Technology, Dec. 1997.
[36] J. Kessens & M. Wester (1997) Improving Recognition Performance by Modelling Pronunciation Variation. Proceedings of the CLS opening Academic Year '97-'98, pp. 1-20. (http://lands.let.kun.nl/literature/kessens.1997.1.html)
[37] A. Kipp, M.-B. Wesenick & F. Schiel (1996). Automatic detection and segmentation of pronunciation variants in German speech corpora. Proc. of ICSLP-96, Philadelphia, pp. 106-109.
[38] A. Kipp, M.-B. Wesenick & F. Schiel (1997). Pronunciation Modeling Applied to Automatic Segmentation of Spontaneous Speech. Proc. of EuroSpeech-97, Rhodes, Vol. 2, pp. 1023-1026.
[39] L. Pousse & G. Perennou (1997) Dealing with pronunciation variants at the language model level for automatic continuous speech recognition of French. Proc. of Eurospeech-97, Rhodes, Vol. 5, pp. 2727-2730.
[40] H. Strik (1998) Publications on pronunciation variation and ASR. http://lands.let.kun.nl/Tspublic/strik/pron-var/references.html
[41] D. Torre, L. Villarrubia, L. Hernandez & J.M. Elvira (1997) Automatic Alternative Transcription Generation and Vocabulary Selection for Flexible Word Recognizers. Proc. of ICASSP-97, Munich, Vol. 2, pp. 1463-1466.
[42] M.-B. Wesenick (1996) Automatic generation of German pronunciation variants. Proc. of ICSLP-96, Philadelphia, pp. 125-128.