

Detailed pronunciation variant modeling for speech transcription

Denis Jouvét, Dominique Fohr, Irina Illina

Speech Group, LORIA-INRIA, 615 rue du Jardin Botanique, 54602 Villers les Nancy, France

{denis.jouvet, dominique.fohr, irina.illina}@loria.fr

Abstract

Modeling pronunciation variants is an important topic for automatic speech recognition. This paper investigates the pronunciation modeling at the lexical level, and presents a detailed modeling of the probabilities of the pronunciation variants. The approach is evaluated on the French ESTER2 corpus, and a significant word error rate reduction is achieved through the use of context and speaking rate dependent modeling of these pronunciation probabilities. A rule-based approach makes it possible to derive a priori probabilities for the pronunciation of words that are not present in the training corpus, and a MAP estimation process yields reliable estimates of the pronunciation variant probabilities.

Index Terms: speech recognition, pronunciation variants, lexical modeling, speaking rate dependent modeling.

1. Introduction

Modeling pronunciation variation is an important topic for automatic speech recognition [1]. It has been widely observed that speech recognition performance degrades notably on spontaneous speech, and more precisely, that the word error rate (WER) increases when the degree of spontaneity increases [2]. The rate of speech is also an important variability source which impacts notably on the acoustic realization of the sounds as well as on the pronunciation of the words [3], and consequently affects recognition performance. Large increases in WER are observed when speaking rate increases [4]. It should be noted that rate of speech and spontaneous speech are not completely independent as the rate of speech is an important cue for detecting spontaneous speech [2][5].

Pronunciation variation impacts on the acoustic, lexical and language modeling levels [1]. Speaking rate influence has been investigated using several sets of models (for example slow vs. fast speech) used either for rescoring or in new decoding after estimation of the current speaking rate. In [6], a set of parallel rate-specific acoustic and pronunciation models are used to represent slow and fast speech; and some pronunciation variants are handled through "zero length" phones, to deal with phone deletions observed in fast speaking rate. At the language modeling level, language models adapted on spontaneous speech have also been investigated [2].

A more general framework for handling speech variability at the acoustic and lexical modeling levels was proposed in [7] through hidden variables that take into account the so called speaking mode. Such framework avoids a hard decision on the variability or speaking mode estimate. The underlying idea is to reduce the confusability by conditioning the modeling on hidden variables, that are estimated either directly during decoding, or from a previous decoding pass.

Large vocabulary speech recognition systems currently use word pronunciation probabilities. However, as training sets are limited, many words in the lexicon are not present in the training set, hence the probability of the corresponding pronunciation variants cannot be estimated. Nevertheless, as

the pronunciation variants are more important for frequent words [4], the corresponding pronunciation variant probabilities are rather correctly estimated for those words. Rules have also been used for obtaining spontaneous speech pronunciation variants [5], and the pronunciation variant probabilities were derived from the probabilities of the rules.

In French, the optional schwa (/ə/) and the liaisons are important pronunciation variants [8]. The optional schwa leads to pronunciation variants that are frequent. The optional schwa can be internal to a word, as for example in the word *semaine* ("week") which can be pronounced /səməɛn/ or /smɛn/; or final to a word, as in the word *le* ("the") which can be pronounced /lə/ or /l/. In this last case, when the schwa is omitted, this can lead to confusion with the reduced form of determiners and pronouns used before a word starting with a vowel (for example *l'* - reduced form of *le/la* ("the") - and pronounced /l/). The liaison is another complicated phenomenon which consists in the realization of a normally mute final consonant when the following word begins with a graphemic vowel or a mute h [8], for example *les états* ("the states") pronounced /lezeta/ where the consonant /z/ is added at the end of word *les* /le/ making the liaison with the following word. When the following word starts with a consonant there is no liaison, as for example *les pays* ("the countries") which is pronounced /lepɛi/. The most common liaison consonants are /z/ and /t/.

This paper focuses on the lexical level modeling, and evaluations are conducted on the French ESTER2 speech corpus [9]. The proposed modeling of the pronunciation variant probabilities, which takes into account a dependence on the following word, as well as on the speaking rate, is described in Section 2, along with the estimation of these probabilities based on rules and on a MAP process. Section 3 presents and discusses the recognition experiments. Finally conclusions are drawn in Section 4.

2. Variability dependent modeling

2.1. Mathematical framework

In the current standard modeling approach, the decoding of an acoustic observation $x = (x_1, \dots, x_T)$ consists in finding the most likely sequence \hat{w} knowing the observation:

$$\begin{aligned} \hat{w} &= \operatorname{argmax}_w p(w|x) \\ &= \operatorname{argmax}_w p(x|w)p(w) \\ &\approx \operatorname{argmax}_{w,q} p(x|q)p(q|w)p(w) \end{aligned} \quad (1)$$

where $w = (w_1, \dots, w_N)$ is a sequence of words, and q is a sequence of phones corresponding to a pronunciation of the word sequence w .

Taking into account variability dependent modeling, using variability variable v estimated from cues y , as done in [7], the decoding equation becomes:

$$w = \operatorname{argmax}_{w,q} \sum_v p(x|q,v)p(q|w,v)p(w|v)p(v|y) \quad (2)$$

where \mathbf{v} represents the variability sources taken into account, such as spontaneous vs. prepared speech, speaking rate, etc.

The paper focuses on the second term of the product, that is the dependency of the pronunciation variant probability on the sequence of words \mathbf{w} and the variability variable \mathbf{v} . In the reported experiments, the variability under consideration is the speaking rate, and the acoustic and language models are independent of the speaking rate. Thus the decoding equation used is:

$$\mathbf{w} = \underset{\mathbf{w}, \mathbf{q}}{\operatorname{argmax}} \sum_{\mathbf{v}} p(\mathbf{x}|\mathbf{q})p(\mathbf{q}|\mathbf{w}, \mathbf{v})p(\mathbf{w})p(\mathbf{v}|\mathbf{y}) \quad (3)$$

As the speaking rate (variability variable \mathbf{v}) is estimated in a deterministic way from the cues \mathbf{y} , the equation becomes:

$$\mathbf{w} = \underset{\mathbf{w}, \mathbf{q}}{\operatorname{argmax}} p(\mathbf{x}|\mathbf{q})p(\mathbf{q}|\mathbf{w}, \mathbf{v})p(\mathbf{w}) \quad (4)$$

The phone sequence \mathbf{q} is the concatenation of the pronunciation variant q_i of each word w_i of the word sequence \mathbf{w} . Hence the probability of the phone sequence \mathbf{q} for the current phrase (speech segment on which the decoding is applied), is obtained by:

$$p(\mathbf{q}|\mathbf{w}, \mathbf{v}) = \prod_i p(q_i|\mathbf{w}, \mathbf{v}) \quad (5)$$

which clearly express the dependency of the probability of the pronunciation variant q_i of each word w_i on the whole word sequence \mathbf{w} , and on the variability variable \mathbf{v} .

2.2. Standard modeling

Baseline modeling: The baseline processing consists in ignoring the probability of the pronunciation variants of each word, this amounts to considering that

$$p(q_i|\mathbf{w}, \mathbf{v}) \triangleq 1.0 \quad (6)$$

which, actually, is not a probability, as this does not sum to 1.0 for words that have several pronunciation variants, but this corresponds to the actual computations carried out in the baseline decoder.

Uniform pronunciation variant probabilities: When no other information is available, another alternative consists in using uniform probabilities for the pronunciation variants of any given word, that is:

$$p(q_i|\mathbf{w}, \mathbf{v}) \triangleq p(q_i|w_i) = \frac{1.0}{\text{Nb variants } w_i} \quad (7)$$

where q_i is a pronunciation variant of word w_i .

2.3. Word context and liaisons

In French some liaisons are compulsory, some are prohibited, and some other are optional. However, a corpus-based study [10] showed that compulsory liaisons are not always realized and that prohibited liaisons are sometimes realized. Nevertheless the liaisons concern sequences of two words, in which the second one begins with a vowel; liaisons consonants are not realized before a word beginning with a consonant.

In a similar way, reduced form of determiners and pronouns (such as l' , d' , s') are used only before words beginning with a vowel.

Penalty for impossible sequences: As mentioned before, some pronunciation variants are impossible when the following word does not begin with a vowel. Those impossible sequences can be handle in a rough (but efficient) way by setting a large penalty on such sequences in the decoding pass. This amounts to:

$$p(q_i|\mathbf{w}, \mathbf{v}) \triangleq p(q_i|w_i, w_{i+1}, \mathbf{v}) = p(q_i|w_i, \mathbf{v}) \cdot \delta(w_i, w_{i+1}) \quad (8)$$

where q_i is a pronunciation variant of the word w_i ; and w_{i+1} is the word following w_i , and

$$\delta(w_i, w_{i+1}) = \begin{cases} 10^{-\text{penalty}} & \text{if impossible sequence} \\ 1.0 & \text{otherwise} \end{cases} \quad (9)$$

Such penalty is easily handled in the backward pass of a forward/backward decoder, as for example in the Julius decoder [11] used in the current experiments.

Word context dependent modeling: Impossible sequences are associated to a following word that does not begin with a vowel, whereas liaisons are associated to sequences in which the following word begins with a vowel. Hence it seems relevant to consider two contexts in the modeling, one C_{+V} corresponding to the following words w_{i+1} that are acceptable in liaisons (roughly French words beginning by a vowel sound), and another one C_{-V} for the other words. This leads to:

$$p(q_i|\mathbf{w}, \mathbf{v}) \triangleq p(q_i|w_i, C_{i+1}, \mathbf{v}) \quad (10)$$

where C_{i+1} equals C_{+V} or C_{-V} depending on word w_{i+1} .

2.4. Speaking rate dependent modeling

Short pronunciation variants (in number of phonemes) are often favored at rapid speaking rate, whereas long ones are more frequently observed in slow speaking rate. One of the most striking example relates to the French schwa which can be pronounced or not, and is illustrated in Figure 1.

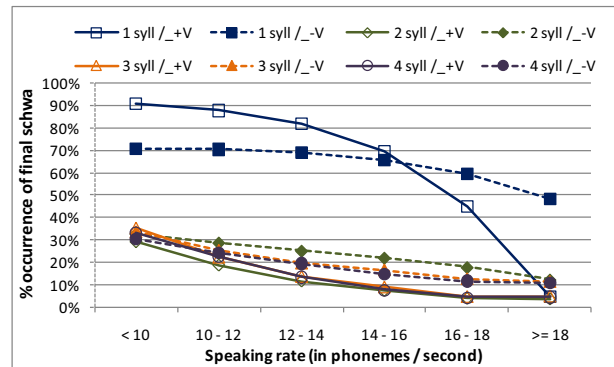


Figure 1: Frequency of occurrence of the final schwa with respect to speaking rate for words of different lengths.

The figure clearly shows that the frequency of occurrence of the final schwa gets lower as the speaking rate increases. Also, the frequency differs, depending on the following word - beginning with a consonant ($-+V$) or not ($-V$). At high speaking rates, the final schwa is more frequently omitted when the following word begins with a vowel (solid lines), than before a word beginning with a consonant (dashed lines). This lead to investigating a speaking rate dependent (SRD) modeling of the probabilities:

$$p(q_i|\mathbf{w}, \mathbf{v}) \triangleq p(q_i|\mathbf{w}, v_k) \quad (11)$$

where v_k is the quantized speaking rate of the speech segment to be recognized. In the following experiments, the speaking rate is estimated from the result of the forward pass of the decoder. It should be noted that here the cues \mathbf{y} are the word segmentations provided by the forward pass of the decoder.

2.5. Estimating pronunciation variant probabilities

Estimating probabilities of pronunciation variants from their frequency of occurrences in a training set provides good estimates for frequent words, but no estimate at all for unseen words and unreliable estimate for words that occur just a few times. Thus the following estimation procedures were investigated and compared.

Frequency of occurrences for frequent words: In this approach the probability of the pronunciation variants of frequent words are the frequency of occurrences of the given pronunciation variants in the training set. Obviously, the frequencies are computed according to the modeling conditioning, that is with or without taking into account the context of the following word, and/or the speaking rate.

For other words, the probabilities of the pronunciation variants are set uniform (i.e. set to $1.0/nb_variants$). In the following experiments, words observed more than 100 times in the training set were considered as frequent ones.

Rule-based a priori and MAP estimation: In order to provide an estimate for the pronunciation variants of unseen words, rules are defined and used; they refer for example to the presence ($\text{ə}<n>+$) or absence ($\text{ə}<n>-$) of the schwa ($/\text{ə}/$) in the n^{th} syllable of a word, or in final position ($\text{ə}\$+$ vs. $\text{ə}\$-$), or the presence ($z\$+$) or absence ($z\$-$) of a liaison consonant (here $/z/$) at the end of a word.

As the frequency of occurrences of the final schwa vary according to the length of the words (see Figure 1 - very different behavior for 1 syllable words vs. 2 or more syllable words), the rule probabilities are computed according to the speaking rate ($v_k; k = 1..K$), to the context (C_{+v} or C_{-v}), and to the number of syllables of the words. A Bayesian MAP estimation process is used which takes into account the actual counts observed in the training data and an a priori that correspond to the probabilities of the generic rules (that is ignoring the number of syllables of the words).

The rule probabilities are then used to compute an a priori probability for the pronunciation variants for which rules apply; otherwise the a priori probabilities are set uniform. Then, using these a priori estimates and the actual counts obtained from the training set, a Bayesian estimation is made for the probability of each word pronunciation variant.

3. Experiments

3.1. Experimental setting

The speech recognition experiments were conducted using the ESTER2 speech corpus [9] which consists of French broadcast news. The data from the African radios were not used here. The number of occurrences of the pronunciation variants was counted on the training set (about 180 hours of signal and 2 M running words). More precisely, the number of occurrences of the pronunciation variants have been counted on the correctly recognized words of the training set.

The impact of various modeling parameters was investigated and analyzed on a large subset of the development set (about 4h30 of signal and 36800 running words). Finally a few modeling configurations were also evaluated on the test set (about 5h50 of signal and 63000 running words).

The experiments were conducted using the ANTS speech transcription system [12], which is based on the HTK toolkit [13] and the Julius decoder [14][11], and now includes an HLDA transform. The lexicon contains about 63000 words. The pronunciation probabilities are affected by the same fudge factor as the language model probabilities. For speaking rate dependent models (SRD models), the speaking rate for each segment is computed from the result of the forward pass decoding, and used in the backward pass of the Julius decoder.

3.2. Word context and liaisons

The first set of experiments involves the baseline models, and probabilities directly estimated from frequency of occurrences of the pronunciation variants.

Table 1. WER on development set, according to the handling of word context and liaisons.

		Penalty imp. seq.	
		Without	With
---	$p(q_i w_i) \triangleq 1.$	27.15%	26.71%
---	$p(q_i w_i) \triangleq 1/NbVar$	26.95%	26.53%
No context	Frequency occurrences	26.22%	26.01%
C_{+v} vs. C_{-v}	Frequency occurrences	26.11%	25.92%

Table 1 shows that using uniform probabilities ($1/NbVar$) leads to better recognition performance than ignoring them ($p(q_i|w_i) \triangleq 1.0$). A large improvement is obtained with the introduction of the probability of the pronunciation variants, estimated from their frequency of occurrences for the frequent words (and uniform probabilities for unseen and less frequent words). The last line shows that counting the frequency of occurrences with respect to the context of the following word provides a further improvement.

The last column shows that, in all cases, introducing a penalty for impossible pronunciation variants sequences provides a noticeable reduction in WER, and also nicely combines with the context dependent modeling.

3.3. Estimation of pronunciation probabilities

The second set of experiments concerns the usage of the MAP estimation procedure, which takes into account both an a priori probability (computed from rules that apply) and the actual frequency of occurrences measured on the training set. When no rules apply, uniform probabilities are used. Two cases are considered for the rules: either only the rules associated to the schwa ($/\text{ə}/$) or all the rules (schwa & liaisons).

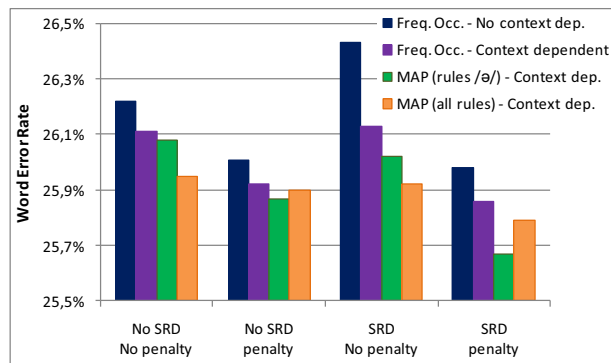


Figure 2: Impact of MAP estimation on WER on development set, using context-dependent probabilities for pronunciation variants, without and with speaking rate dependency (SRD).

Figure 2 shows that in all cases MAP estimates lead to better performance than the frequency of occurrences. The third set of bars (green bars - i.e. using only the rules associated to the schwa) shows that the improvement is larger for the speaking rate dependent (SRD) models. These models have more parameters (the amount is proportional to the number of bins for quantizing the speech rate), hence the larger benefit of applying rules and MAP estimation.

The last set of bars (orange bars - i.e. using all rules) shows a further improvement when using also the rules for the liaisons but only when no penalty is applied on the impossible pronunciation variant sequences. When a penalty is used for impossible sequences, the adjunction of the rules associated to the liaisons does not bring any improvement. This could be explained by a redundant handling of the liaisons in this case - penalty on one side; and rules on the other side - or liaison rules that are not adequate enough.

3.4. Context and speaking rate dependent modeling

Figure 2 clearly shows the benefit of context dependent modeling in all conditions (i.e. with or without penalty and with or without speaking rate dependent - SRD - modeling).

In the reported experiments, the speaking rate was quantized in 6 bins, as indicated in Figure 1 (horizontal axis). As this increases the amount of parameters (one set for each speaking rate bin), there is no improvement observed when probabilities of pronunciation variants are directly approximated by the counts observed in the training data. However, as Figure 2 shows, when a MAP estimation of the pronunciation variant probabilities is used, the detailed modeling that depends on the following word context and on the speaking rate provides an improvement.

3.5. Evaluation on test set and discussion

Table 2 reports the results (and confidence intervals) of the evaluation of some of the previous configurations on the test set. The behavior observed on the test set is very similar to the one observed on the development set, and this confirms that it is important to use adequate probabilities for the pronunciation variants. Moreover, the results shows the benefit of a detailed modeling of the probabilities of the pronunciation variants, which takes into account the dependency on the context and on the speaking rate, together with a MAP and rule based estimation.

Table 2. WER on test set for various modeling of the pronunciation variant probabilities.

	WER
$p(q_i w_i) \triangleq 1$. (no penalty)	26.39% ($\pm 0.34\%$)
$p(q_i w_i) \triangleq 1/\text{NbVar}$ (no penalty)	26.03% ($\pm 0.34\%$)
Freq. occ., (no context, no penalty, no SRD)	25.88% ($\pm 0.34\%$)
MAP (rules \varnothing , C_{+V} vs. C_{-V} , penalty, SRD)	25.29% ($\pm 0.34\%$)

In the current experiments, the rules associated to the liaisons did not take into account any morpho-syntactic information. As such information is relevant (e.g. [10]), we could expect that further improvement could be achieved with a refined set of rules for the estimation of a priori probabilities for the liaisons. For what concerns the dependency of the liaisons and schwa on the frequency of the corresponding words [15], this is taken into account by the MAP estimation procedure. The rules are used only for computing a priori estimates, and the MAP procedure makes a compromise between these a priori estimate and the actual frequencies from the training set.

Because of the strong link between the speaking rate and the spontaneity of the speech signal [2][5], the current speaking rate dependent modeling that was proposed here handle a part of the variability associated to spontaneous speech, especially for what concerns the pronunciation, or omission, of the schwa, which is an important phenomenon of spontaneous speech [16].

4. Conclusions

In this paper we have analyzed in details the modeling of the probabilities of the pronunciation variants. Although this represents a very small amount of parameters with respect to the acoustic model parameters and language model parameters, a detailed modeling of those probabilities leads to a significant improvement in the word error rate, as measured here for French broadcast news transcription.

It was shown that it is important to take into account in the modeling, factors that impact on the pronunciation variants, as for example the dependency on the context of the following

word (starting with a vowel or not) as well as the dependency on the speaking rate. Such modeling nicely combines with the usage a penalty for forbidding impossible pronunciation sequences. Moreover the usage of rules makes it possible to estimate a priori probabilities for pronunciation variants, which are then used in the final MAP estimation process.

Further work will concern the refinement of the rules, especially for the liaisons by taking into account morpho-syntactic information, and also the extension of the set of rules for handling other spontaneous speech specific pronunciation variants; as well as extending the speaking mode dependency to the acoustic and language models.

5. References

- [1] Strik, H., and Cucchiari, C., "Modeling pronunciation variation for ASR: A survey of the literature", *Speech Communication* 29, pp. 225-246, 1999.
- [2] Dufour, R., Jousse, V., Estève, Y., Béchet, F., and Linares, G., "Spontaneous speech characterization and detection in large audio database", *Proc. Specom'2009, Int. Conf. on Speech and Computer*, St. Petersburg, Russia, pp 41-46, 2009.
- [3] Benzeghiba, M., de Mori, R., Deroo, O., Dupont, S., Erbes, T., Jouviet, D., Fissore, L., Laface, P., Mertins, A., Ris, C., Rose, R., Tyagi, V., and Wellekens, C., "Automatic speech recognition and variability: a review", *Speech Communication* 49, pp. 763-786, 2007.
- [4] Fosler-Lussier, E., and Lorgan, N., "Effects of speaking rate and word frequency on pronunciations in conversational speech", *Speech Communication* 29, pp. 137-158, 1999.
- [5] Finke, M., and Waibel, A., "Speaking mode dependent pronunciation modeling in large vocabulary conversational speech recognition", *Proc. EUROSPEECH'97*, Rhodes, Greece, pp. 2379-2382, 1997.
- [6] Zheng, J., Franco, H., and Stolcke, A., "Modeling word-level rate-of-speech variation in large vocabulary conversational speech recognition", *Speech Communication* 41, pp. 273-285, 2003.
- [7] Ostendorf, M., Byrne, B., Bacchiani, M., Finke M., Gunawardana, A., Ross, K., Roweis, S., Shriberg, E., Talkin, D., Waibel, A., Wheatley, B., and Zeppenfeld, T., "Modeling Systematic Variations in Pronunciation via a Language-Dependent Hidden Speaking Mode", Hopkins University, 1996.
- [8] Adda-Decker, M., Boula de Mareuil, P., and Lamel, L., "Pronunciation variants in French: schwa & liaison", *Proc. ICPhS'99*, San Francisco, USA, pp. 2239-2242, 1999.
- [9] Galliano, S., Gravier, G., and Chaubard, L., "The Ester 2 evaluation campaign for rich transcription of French broadcasts", *Proc. INTERSPEECH'2009*, Brighton, UK, pp. 2583-2586, 2009.
- [10] Boula de Mareuil, P., Adda-Decker, M., and Gendner, V., "Liaisons in French: a corpus-based study using morpho-syntactic information", *Proc. ICPhS'2003*, Barcelona, Spain, pp. 1329-1332, 2003.
- [11] Lee, A. and Kawahara, T., "Recent Development of Open-Source Speech Recognition Engine Julius", *Proc. APSIPA ASC'2009, Asia-Pacific Signal and Information Processing Association Annual Summit and Conf.*, Sapporo, Japan 2009.
- [12] Illina, I., Fohr, D., Mella, O., and Cerisara C., "The Automatic News Transcription System: ANTS - some Real Time experiments", *Proc. ICSLP'2004*, Jeju Island, Korea, pp. 377-380, 2004.
- [13] <http://htk.eng.cam.ac.uk/>
- [14] http://julius.sourceforge.jp/en_index.php
- [15] Fougeron, C., Goldman, J.P., and Frauenfelder, U.H., "Liaison and schwa in French: an effect of lexical frequency and competition?", *Proc. EUROSPEECH'2001*, Aalborg, Denmark, pp. 639-642, 2001.
- [16] Bazillon, T., Jousse, V., Béchet, F., Estève, Y., Linares, G., and Luzzati, D., "La parole spontanée : transcription et traitement", *Traitement Automatique des Langues* 49(3), pp 47-76, 2008.