BAYESIAN FRAMEWORK FOR VOICING ALTERNATION & ASSIMILATION STUDIES ON LARGE CORPORA IN FRENCH

Martine Adda-Decker[1]*, Pierre Hallé*[2]

[1]LIMSI/CNRS, 91403 Orsay, [2]LPP/CNRS 19 rue des Bernardins, 75005 Paris
madda@limsi.fr,pierre.halle@univ-paris5.fr

ABSTRACT

The presented work aims at exploring voicing alternation and assimilation on very large corpora using a Bayesian framework. A voice feature (VF) variable has been introduced whose value is determined using statistical acoustic phoneme models, corresponding to 3-state Gaussian mixture Hidden Markov Models. For all relevant consonants, i.e. oral plosives and fricatives their surface form voice feature is determined by maximising the acoustic likelihood of the competing phoneme models. A voicing alternation (VA) measure counts the number of changes between underlying and surface form voice features. Using a corpus of 90h of French journalistic speech, an overall voicing alternation rate of 2.7% has been measured, thus calibrating the method's accuracy. The VA rate remains below 2% word-internally and on word starts and raises up to 9% on lexical word endings. In assimilation contexts rates grow significantly ($> 20\%$), highlighting regressive voicing assimilation. Results also exhibit a weak tendency for progressive devoicing.

Keywords: voicing assimilation, voicing alternation, automatic speech alignment, Bayesian decision

1. INTRODUCTION

The progress achieved these last decades in automatic speech recognition is largely due to statistical modeling in a Bayesian decision framework [6]. This framework can be adapted for linguistic studies and speech recognizers may be tuned to provide large-scale acoustic-phonetic descriptions which correspond to surface forms of underlying phonemic baseforms [5, 1]. In the present contribution we are interested in voicing alternation which consists for a given consonant with underlying voice feature f, alternation of its surface form to voice feature $\neg f$. More particularly, we want to examine voicing assimilation which is known to be a major factor of voicing alternation in French [7, 9, 3]. Hence surface forms must allow alternation between a voiced consonant and its voiceless counterpart and vice-versa. For example the word *chef* with baseform pronunciation /ʃɛf/ must include [ʃɛf], [ʒɛf], [ʃɛv] and [ʒɛv] in the set of possible surface forms. As voicing assimilation in French is known to be regressive [7] with voicing alternation on word endings, the alternate form [ʃɛv] is expected to be the most promising variant. Typical examples of potential assimilation contexts in French involve frequent function word sequences (*avec des, plus de*), noun preposition (*politique de, processus de, chef de...*), involving *de* (the most frequent function word in French), article noun (*cette guerre*), verb conjunction (*trouve que), verb adverb (*peuvent pas), noun adjective (*république démocratique, chaque jour, étape judiciaire*), proper names (*Yves Saint Laurent, Dominique Voynet*), as well as dates and numeral-noun combinations (*sept décembre, quinze chars*). Many of the preceding examples such as *chaque jour* (*each day*), *cette guerre* (*this war*) may be produced with a linking schwa thus reducing the assimilation potential. This also holds for word sequences spanning across phrase boundaries, where pauses or respirations may be inserted.

In the present study the addressed questions are the following: on a methodological level we examine whether the proposed method is sound for large-scale voicing alternation and assimilation studies. To what extend the observed results can be considered as reliable? On a more linguistic level we examine voicing alternation tendencies with respect to different conditions: complete corpus versus function words and lexical words; position of the consonant in the word; different assimilation contexts. The structure of the paper is as follows. First we present the methodology to explore voicing alternation. Section 3. provides a description of the corpus, including frequency counts of potential assimilation contexts. Section 4. details results concerning voicing alternation and assimilation in different conditions. Section 5. summarizes the discussion.

2. METHOD

2.1. Voicing alternation

In French voicing is distinctive for oral plosives (/p/, /t/, /k/, /b/, /d/, /g/) and for fricatives (/f/, /s/, /ʃ/, /v/, /z/, /ʒ/). We define **voicing alternation** for these

phonemes as a change of the underlying voicing feature f to its opposite value $\neg f$ in its surface form.

2.2. Voicing assimilation

Voicing assimilation can be seen as a particular case of voicing alternation, with constraints on contexts where voicing alternation is allowed. Voicing assimilation consists for a given consonant in inheriting the voicing feature (voiced, voiceless) of an adjacent consonant. The phenomenon is considered here as categorical even though partial assimilation can sometimes be observed [8]. In French voicing assimilation is known to be regressive, i.e. the voicing feature is inherited from the following consonant (e.g. *sub-saharien*: the /b/ of the prefix *sub* is most probably realised as [p] due to the adjacent voiceless fricative /s/ of the word start *saharien*).

2.3. Bayesian framework

For each relevant consonant type φ voicing alternation can be described as a classification problem with a finite set of 2 states of possible voicing features $VF = \{V, NV\}$, V corresponding to the *voiced* state, NV to the *not voiced* or *voiceless* state. Bayes decision rule gives:

$$v^* = \arg\max_{v \in VF} P(\varphi_v|x) \quad (1)$$

$$= \arg\max_{v \in VF} P(\varphi_v) p(x|\lambda_{\varphi_v}) \quad (2)$$

with $v \in VF, P(\varphi_v)$ the prior probabilities, λ_{φ_v} the acoustic phone HMM model and $p(x|\lambda_{\varphi_v})$ the conditional probability density functions. Prior probabilities allow to optimise for the classification of the underlying baseform feature. As we are interested in the surface form, modeled by the conditionals $p(x|\lambda_{\varphi_v})$, priors are set equal ($P(\varphi_v) = P(V) = P(NV)$), which simplifies to:

$$v^* = \arg\max_{v \in VF} p(x|\lambda_{\varphi_v}) \quad (3)$$

The voicing decision hence fully relies on the conditional probability densities, described by 3-state Gaussian mixture HMM models (256 G/state).

2.4. Speech alignment

The above described V-NV classification problem is addressed during speech alignment using context-independent acoustic phone models and voicing alternation specific variants (see below). Context-dependent models are known to be less accurate than context-independent models for automatic phonetic segmentation [1]. The acoustic models used here have been trained on a subset of journalistic speech data including only segments of length greater than 50ms. Shorter segments are known to be "corrupted" by coarticulation effects. Segmentation accuracy might nonetheless be a limiting factor for automated studies, as boundary location is optimized globally with at best a 10 ms precision.

2.5. Voicing alternation specific variants

To focus on voicing feature alternates, surface forms are allowed to alternate from a voiced consonant to its voiceless counterpart and vice-versa. For example the word *chef* with baseform pronunciation /ʃɛf/ includes [ʃɛf], [ʒɛf], [ʃɛv] and [ʒɛv] in the set of possible surface forms. Voicing assimilation being mainly regressive in French with voicing alternation on word endings, the alternate form [ʃɛv] is expected to be the most promising variant.

2.6. Assimilation contexts

For the present study we have defined assimilation contexts as occuring on C1#C2 word boundaries with opposite voicing features between C1 and C2.

2.7. Control conditions

A first control condition has been designed to evaluate the instrument's accuracy with respect to the V/NV decision. This condition consists in measuring V/NV alternation rates (e.g. /b/ recognised as [p]) for all **alternating consonants** (i.e. consonants relevant to voicing alternation studies: /p/, /t/, /k/, /b, /d/, /g/, /f/, /s/, /ʃ/, /v/, /z/, /ʒ/, **independently of the context of occurrence** over the whole corpus (see section 4.1.). Low alternation rates correlate with high V/NV decision accuracy. A second control condition aims at comparing measurements from V#NV and NV#V assimilating contexts to similar non-assimilating V#V and NV#NV conditions.

3. CORPUS

We made use of 90 hours of French broadcast news from the ESTER corpus of the TECHNOLANGUE rich transcription evaluation [2]. Speaking style corresponds to fluent, rather clearly articulated speech with a tendency of stressing word-initial syllables, considered as characterising journalistic speech.

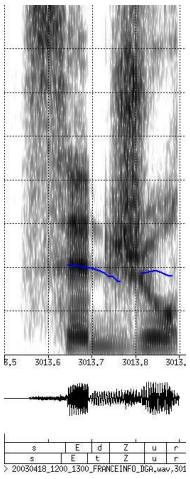
3.1. Frequency of assimilation contexts

Before starting the investigation proper, a preliminary question of interest concerns the frequency of assimilation contexts in running speech. Among a total of 1M word boundaries in the corpus, about 40k word boundaries correspond either to assimilation or to corresponding control contexts (see Table 1). Most frequently observed assimilating contexts are either proper names *Mohamed Six* (174

occ.), *Côte d'Azur* (90 occ.), and sequences involving function words (*de, des, du, d'*) such as *chef de, force de, tête de*, with tens of occurrences each. Frequent POS combinations, such as noun + adjective, tend to have sparse representatives (*trêve conditionnelle, sept dirigeants, roches différentes*). Slightly more word boundaries (17k+6k occurrences) are found for the control condition with stable underlying +/- voice features on the C1#C2 sequence (e.g. *George Bush, Jacques Chirac*).

Table 1: left: Nb. of assimilation, control contexts. **right:** Spectrogram of NV#V assimilation.

word boundaries	
cond	#occ
all	1M
assimil	
NV#V	13k
V#NV	5k
control	
NV#NV	17k
V#V	6k



4. EXPERIMENTAL RESULTS

A first question concerns the rate of alternate alignments independently of assimilation contexts. This question is related to the instrument's accuracy: in non-assimilation contexts, the use of alternates should remain low. In particular word-internally where the rate of assimilation contexts is almost negligible, alternation should remain very low. Alternations on word endings are expected to be higher than on word starts, due to regressive assimilation.

4.1. Overall voicing alternation rates

A voicing alternation (VA) measure counts the number of changes between underlying and surface form voice features on alternating consonants. A VA rate is then defined on alternating consonants (independently of their context) as the number of times an alternating consonant with voice feature f is classified as voice feature $\neg f$. The overall VA rate includes alternations in assimilation contexts, however their proportion among alternating consonants remains low: less than 2% (notice that only half of these are likely to assimilate to the voicing feature of their neighboring half).

A global VA rate of 2.7% is measured on a total of 1M alternating consonants. This low alternation rate gives an idea of the instrument's accuracy, as well as of the methodological validity. Alternations can be partly explained by decision errors, partly by "errors" in the observed speech signal (unexpected variants, overlapping music), partly by assimilations. Alternation rates are next examined according to the position of the consonant in the word skeleton: word-internal vs boundary (for the latter word-start vs word-end). Table 2 shows corresponding alternation figures and frequency counts. Each cell contains a VA rate, together with the corresponding absolute frequency counts (#) from the corpus: $\#(\neg f)$ the number of classifications, where surface and underlying voicing features disagree, and below $\#(f+\neg f)$ the total number of occurrences per condition. Counts are given in k (thousands).

Table 2: VA rates and frequency counts (#) in k(*1000), alternations, total below, for different conditions: overall, word-internal, word-boundary, for the latter word-start vs word-final. 2nd/3rd lines correspond to function/lexical word subsets. The 4th line **All2** includes all but assimilating contexts.

voicing alternation rates					
	overall %VA #	internal %VA #	boundary %VA #	w-start %VA #	w-final %VA #
All	**2.7** 28.7 1060	**1.7** 7.7 449	**3.4** 21.0 611	**2.0** 9.9 461	**7.6** 11.9 150
Fct	**2.7** 7.5 276	**1.3** 0.3 22.7	**2.9** 7.2 253	**2.1** 4.7 202	**5.4** 3.2 51
Lex	**2.7** 21.2 784	**1.7** 7.4 427	**3.8** 13.7 357	**2.0** 5.2 262	**8.9** 8.6 97
All2	**2.3** 22 960	**1.7** 7.2 419	**2.7** 15 541	**2.0** 9 453	**5.1** 6 117

The first line shows that VA rates are lower in word-internal position (1.7%) than on word boundaries (3.4%) confirming prior intuition. Similarly the number of measured alternations is significantly higher on word endings (7.6%) than on word starts (2%), which may already indicate regressive assimilation. The next two lines allow to compare VA rates between function and lexical words. Are there important differences here? Overall alternation rates don't vary. However two observations are noteworthy for function words: they are mainly monosyllabic, hence very few within-word consonants. Secondly, word-final consonants have a lower VA rate in comparison to lexical words. As many consonants stem from *liaisons* phenomena (generally VCV context), they might be less prone to voicing alternation. In order to clarify the impact of assimilation contexts, figures of the last line (**All2**) have been obtained without the words in assimilating C1#C2 contexts. The overall error rate is slightly lower (2.3%) in this condition: no differences are measured for word-internal and word-start condition, however the word-end rate very significantly drops to 5.1%, given the small amount of data re-

moved. More detailed VA rate analyses as a function of voice feature f show that **devoicing** VA rates (e.g. /b/ aligned as [p] in 3.2% of overall /b/ occurrences, and 3.8% on word starts) are globally higher than **voicing** VA rates (e.g. /p/ aligned as [b] in 1.7% of overall /p/ occurrences, and 1.5% on word starts). This is true for all oral plosives and fricatives of the alternating phoneme set. Hence, voicing vs devoicing appears not to be symmetric. This observation might be related to properties of the acoustic models: if voiceless consonants tend to be at least partially voiced, the acoustic models implicitly take into account this voicing property and hence may be selected on voiced segments. This point needs further investigations. To give a clearcut answer, beyond hypotheses, we need to confront the alignment results to objective acoustic measures (F0 voicing) [4]. It might also be related to word-initial stress and potential links between stress and devoicing.

4.2. Voicing alternation in assimilation contexts

The data are partitioned into 5 subsets depending on the word boundary type: two assimilating (NV#V, V#NV), two corresponding control (V#V, NV#NV) conditions and a global control condition including all the remaining items. Table 3 shows VA rates for C1 (word-final) and for C2 (word-initial).

Table 3: VA rates for C1 and C2 consonants (final and initial word boundaries) using 5 complementary conditions. The NON condition includes all word boundaries for which C1 and C2 are not both in the set of alternating consonants.

	%VA rate	*cond.*	%VA rate
	C1	C1#C2	C2
control	9	NV#NV	1
assimil	24	NV#V	4.5
assimil	20	V#NV	1
control	4	V#V	3
control	5	NON	2

Concerning the **C1 consonant**, Table 3 shows a strong tendency to regressive assimilation for both NV#V (24%) and V#NV (20%). A slight asymmetry can be observed in favour of NV#V: a voiceless consonant becomes more often voiced due to regressive assimilation than the reciprocal configuration. Figures also exhibit a weak tendency of **C1-voicing**, independently of regressive assimilation: comparing NV#NV alternation rates (9%) to the corresponding V#V rate (4%) shows that a voiceless consonant becomes easier voiced than the opposite. Concerning the **C2 consonant**, VA rates are very low, underlining the stability of word-start consonants. However a cross-condition comparison reveals two weak tendencies: first progressive assimilation for the NV#V condition (4.5%) and second **C2-devoicing** (3% on V#V, 4.5% on NV#V). In word-initial C2 position, a voiced consonant easier changes to its voiceless counterpart than the reverse.

5. CONCLUSIONS & PERSPECTIVES

In this contribution we propose a Bayesian framework to study voicing assimilation and, more generally voicing alternation. Using a corpus of 90h of French speech, an overall voicing alternation rate of 2.7% has been measured, thus calibrating the method's accuracy. The VA rate remains below 2% word-internally and on word starts and raises up to 9% on lexical word endings. In assimilation contexts rates grow significantly ($> 20\%$), highlighting regressive voicing assimilation. Results also suggest weak tendencies for progressive (devoicing) assimilation, as well as C1 voicing and C2 devoicing independently from assimilation contexts. Beyond the results presented here, our study shows that the identity of assimilating phoneme sequences, as well as lexical cooccurrence frequency, POS, duration and stress are interfering factors. For the assimilation issue proper, additional insight can be gained by a more extensive study of the link between voicing alternation and these factors. Concerning methodological aspects, future work includes confronting automatic alignment results to objective acoustic measures.

6. ACKNOWLEDGMENTS

We are grateful to Lori Lamel for her cooperation.

7. REFERENCES

[1] Cucchiarini C., Strik H., 2003. Automatic phonetic transcription: An overview. *Proc. 15th ICPhS* Barcelona, 347-350.

[2] Galliano, S., et al., 2005. The ESTER PhaseII Evaluation Campaign for the Rich Transcription of French Broadcast News. Interspeech, Lisboa, 1149-1152.

[3] Grammont, M. (1939). *Traité de Phonétique.* Paris: Delagrave.

[4] Hallé, P., Adda-Decker, M., 2007. Voicing assimilation in journalistic speech. *Proc. 16th ICPhS*, Saarbrücken.

[5] Hunt, M.J., 2004. Speech recognition, syllabification & statistical phonetics. *Proc. ICSLP*, Jeju, 739-742.

[6] Jelinek, F. 1998. *Statistical Methods for Speech Recognition*, Bradford Books, MIT Press 1998.

[7] Rigault, A. 1967. L'assimilation consonantique de sonorité en français: étude acoustique et perceptuelle [Voice assimilation of consonants in French: An acoustic and perceptual study]. *Proc. 6th ICPhS*, Prague, 763-766.

[8] Snoeren, N. et al. A voice for the voiceless: Production and perception of assimilated stops in French. *Journal of Phonetics*, 34, 241-268.

[9] Tranel, B. 1987. *The sounds of French.* Cambridge: Cambridge University Press.