Pronunciation Variants Across Systems, Languages and Speaking Style

Martine Adda-Decker and Lori Lamel
Spoken Language Processing Group
LIMSI-CNRS, BP 133, 91403 Orsay cedex, FRANCE
{lamel,madda}@limsi.fr
http://www.limsi.fr/TLP

ABSTRACT

This contribution aims at evaluating the use of pronunciation variants across different system configurations, languages and speaking styles. This study is limited to the use of variants during speech alignment, given an orthographic transcription and a phonemically represented lexicon, thus focusing on the modeling abilities of the acoustic word models. Parallel and sequential variants are tested in order to measure the spectral and temporal modeling accuracy. As a preliminary step we investigated the dependance of the aligned variants on the recognizer configuration. A cross-lingual study was carried out for read speech in French and American English using the BREF and the WSJ corpora. A comparison between read and spontaneous speech is presented for French based on alignments from BREF (read) and MASK (spontaneous) data.

INTRODUCTION

Adding pronunciation variants in a recognition system's lexicon provides a means of increasing acoustic word modeling options. The additional variants are intended to improve the decoding accuracy of the recognizer. However, if the types of variants are inappropriate or simply not relevant with respect to the weakness of the recognizer, its overall performance may decrease. How many times were the new pronunciation variants, which were added to solve a given acoustic modeling problem, globally ineffective? While solving the problem for which they were designed, variants often introduce new errors elsewhere, canceling the local benefit: as variants may increase homophone rates they become potential error sources. Variants are thus introduced carefully in our speech recognition systems.

In this contribution we examine the use of pronunciation variants during speech transcription alignment focusing on the appropriateness of the acoustic word models given the observed acoustic data. The use of automatically generated pronunciation variants is investigated along different axes: system configuration, language and speaking style (read or spontaneous). Pronunciation variants are distinguished as sequential or parallel: sequential variants allow some phones to be optional, hence increasing temporal modeling flexibility. Parallel variants allow alternative phones from an a priori defined subset to replace a given phone.

SPEECH CORPORA

Three corpora were used for these experiments. Two are widely-used read speech corpora: BREF in French and WSJ0 in English. The third is a spontaneous speech corpus in French. The BREF [2] corpus contains 66.5k sentences from 120 speakers reading newspaper articles from the *LeMonde* paper (about 120 hours of acoustic data). Although considerably more data are available for American English, we have used a portion of the WSJ0 data [4] from 110 speakers uttering a total of 10k sentences (21 hours of acoustic data). The spontaneous speech data were recorded for the ESPRIT MASK (Multimodal-Multimedia Automated Service Kiosk) task [3]. From these we used 38k sentences from 409 speakers (35 hours of acoustic data). The contents of these corpora are summarized in Table 1.

Corpus	MASK	BREF	WSJ
language	French	French	English
style	spontaneous	read	read
#words(total)	260k	1.1M	180k
#words(distinct)	2k	25k	11k

Table 1: Language, speaking style, total and distinct number of words for each corpus.

Figure 1 shows the cumulative lexical coverage of the speech corpora as a function of the word frequency rank. For MASK (spontaneous task-oriented speech) the 10 most frequent words account for 30% of the corpus, whereas for read newspaper speech in both languages they cover about 20% of the data. The 100 most frequent words cover 80% of the MASK corpus, but slightly less than 50% of BREF and WSJ. While the read newspaper corpus coverage is seen to be close to linear on the logarithmic scale for both French and English, a much stronger slope is observed for the spontaneous MASK data between ranks 10 and 200 due to the domain-specificity of the corpus.

PRONUNCIATION LEXICA

Starting with our standard pronunciation lexica (reference lexica) we have designed augmented pronunciation lexica allowing either for parallel or sequential variation. Our goal

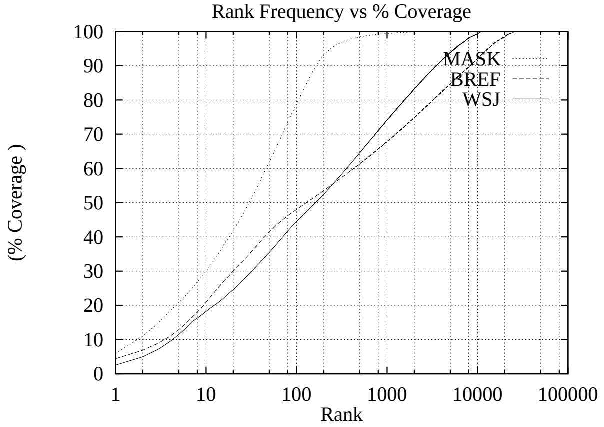

Figure 1: Lexical coverage of spontaneous speech (MASK) and read speech WSJ and BREF corpora.

is to increase our insight in spectral and temporal modeling accuracy and/or weakness in the acoustic models, by comparing alignment results using these different lexica.

Reference lexica

Some example entries from our reference lexica used for training acoustic models are shown in Table 2. These lexica typically contain 10% to 20% pronunciation variants needed to describe alternate pronunciations observed for frequent words (E1 in Tab. 2), proper (particularly foreign) names (E4), for numbers (F4) and acronyms. In French a significant number of variants are introduced to account for word-final optional schwas (F3,F4) and liaisons (F2).

république	repyblik	F1
les	le lez	F2
prendre	prAdr{x} prAd	F3
dix	dis{x} di diz	F4
FOR	fcr fX	E1
THAT	D[@x]t	E2
INVESTMENTS	InvEs{t}mxn{t}s	E3
STEPHEN	stivxn stEfxn	E4

Table 2: Example lexical entries for French (F1-F5) and English (E1-E4) in the reference lexica illustrating parallel ([ ]: alternate phones) and sequential ({ }: optional phones).

Sequential variant lexica

Large sequential variant lexica were automatically derived from the reference lexica by allowing either all vowels or all consonants to be optional. These lexica, *Vopt* and *Copt* (Table 3), aim to locate possible temporal mismatches in the acoustic word models. The *Vopt* lexicon can model the well-known phenomena in French of optional word-final schwas. The *Vopt* lexica can also be used to investigate to what extent and in what contexts non-schwa vowel deletion is observed. Such vowel deletions are usually assumed to be infrequent, but are found in spontaneous speech, entailing syllabic restructuration. In languages with complex consonant clusters, reduction phenomena can be accounted for by introducing sequential variants (E3).

les	l{e} l{e}z
république	r{e}p{y}bl{i}k
FOR	f{c}r f{X}
STEPHEN	st{i}v{x}n st{E}f{x}n

Table 3: Example lexical entries in the *Vopt* lexica illustrating the augmented sequential flexibility.

Parallel variant lexica

Parallel variant lexica have been generated by defining a variety of broad phone classes and allowing each phone in a given class to be replaced by any member of the same class. For each broad phone class a specific lexicon was generated. Table 4 lists the phone classes reported on here.

	French	*English*
Vclass1	Ee	iI|Ye
Vclass2	IxXc	XRx
Cclass1	bdgv	bdgvw
Cclass2	lrhwj	Llryhw

Table 4: Phone classes for the parallel variant lexica design.

In French, many quasi-homophones are separated by the open-closed distinction on vowels (e.g.: `est /E/`, `et /e/`, verbs ending in `-er /e/`, past participle endings `-é /e/`, past tense endings `-ai,ais,ait /E/`). In fluent speech the open-closed distinction may disappear, word identification relying increasingly on higher level constraints (lexical, syntactic, pragmatic, . . .).

Table 5 indicates the complexity of the sequential and parallel variant lexica as the unweighted ratio of the *total number of variants* and the *total number of entries*. The *Copt* lexica has the highest number of variants for all corpora. The *Vopt* lexica contains about half the number of the *Copt* lexica for French, and about one third for English, due to the higher density of consonants in English than in French. The larger figures for BREF may simply be due to the larger lexicon size (cf. Tab. 1), as less frequent words tend to be longer. Since fewer phones can be modified, the parallel variant lexica have a lower complexity, with the largest values for French *Cclass2* (liquids and glides) and the English *Vclass1* (front vowels).

USE OF PRONUNCIATION VARIANTS

We have chosen to measure the use of pronunciation variants by counting the number of word occurrences aligned with alternate pronunciations. In particular we define the *variant2+* rate, which is the percentage of word occurrences aligned with the variants of frequency rank 2 or higher. This measure may be indicative of the possible need for pronunciation variants in the recognition system or equivalently of the appropriateness of a unique acoustic word model as generated by the most frequently used phone transcription.

	MASK	*BREF*	*WSJ*
Reference	1.1	1.2	1.2
Vopt	9.5	17.3	8.2
Copt	20.0	33.7	24.1
Vclass1	1.7	2.5	8.1
Vclass2	2.4	4.0	3.1
Cclass1	2.7	4.3	5.8
Cclass2	10.1	15.1	6.9

Table 5: Unweighted ratios $\frac{\#variants}{\#entries}$ in reference, sequential *Vopt* and *Copt* lexica, and parallel *Vclass* and *Cclass* lexica.

In the following figures, results are displayed as a function of word frequency rank, since the acoustic variability is generally higher for frequent words. The variant2+ rate of the corresponding reference lexicon is included for comparison.

System configuration

To investigate the dependance of the choice of variant on the system configuration, alignment experiments using different acoustic model sets (36, 35, 46 context-independent and 637, 594, 653 context-dependent models, respectively for MASK, BREF and WSJ) have been carried out.

In Table 6 a significant decrease in the *variant2+* rate is observed with context-dependent (CD) models as compared to context-independent models (CI) for alignments using the *Vopt* and *Copt* lexica. Similar *variant2+* rate reductions were observed for all tested lexica. An increasing number of CD acoustic models tends to reduce the need for pronunciation variants, as shown in Figures 2 and 3.

		MASK	*BREF*	*WSJ*
Vopt	*CI*	22.2	18.6	15.7
	CD	13.0	14.8	9.9
Copt	*CI*	27.0	21.0	21.5
	CD	14.5	16.2	12.5

Table 6: Percentage of word occurrences aligned with a phonemic variant of frequency rank 2 or more (*variant2+* rate) for different acoustic model sets and lexica.

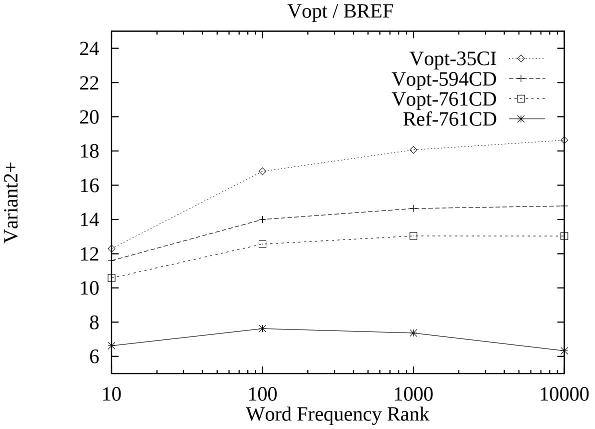

Figure 2: *Variant2+* rate vs Rank for French (BREF) using the *Vopt* lexica and different acoustic model sets.

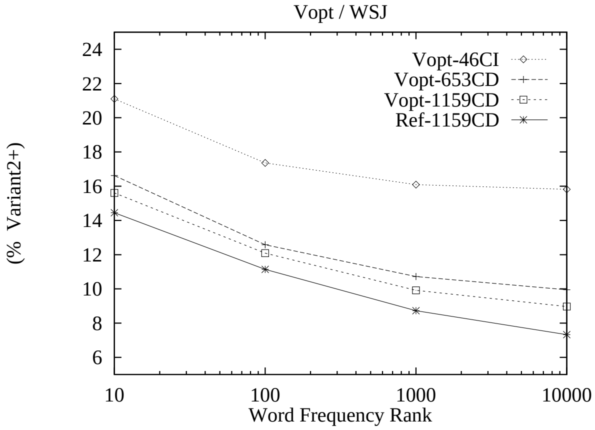

Figure 3: *Variant2+* rate vs Rank for English (WSJ) using the *Vopt* lexica and different acoustic model sets.

Language

In this section we compare the *variant2+* rate for the French and English read speech corpora, using different acoustic model sets and different variant lexica. In Figures 2 and 3 the *Vopt* lexica have been used. Corresponding curves with the *Copt* lexica are shown in Figure 4. For French, the *variant2+* rate increases with frequency rank, whereas the corresponding English curves decrease substantially with frequency rank. The observations for English satisfy our linguistic intuition about acoustic variability and word frequency: the acoustic models seem to accurately represent phones, resulting in a larger variant rate for frequent words. In contrast, for French, the acoustic models seem to well represent the more frequent words, but are less appropriate for infrequent words. A related factor is that there are very few variants for the most common words in the reference lexicon used to train the acoustic models. As expected from a priori linguistic knowledge our measures show a higher *variant2+* rate for French sequential lexica than for English.

Acoustic models for consonants are relatively accurate for French and inversely less discriminative for English (see Figures 5 and 6). The opposite is observed for the vowel classes in the two languages (see Figures 7 and 8). The *Vclass1* in French has a high *variant2+* rate, with a large proportion of E→e substitutions. Despite high complexities in the corresponding lexica, French *Cclass2* and English *Vclass1* obtain low *variant2+* rates.

Read versus spontaneous speech

The Mask *variant2+* rate curves globally decrease, as did the WSJ ones. To understand the difference in behavior of the MASK and BREF data, we looked at the number of variants weighted by their corresponding word frequencies in the training corpus. Using this measure, the curves as a function of word frequency rank are essentially parallel to the *variant2+* curves of the reference lexica. Considering only the

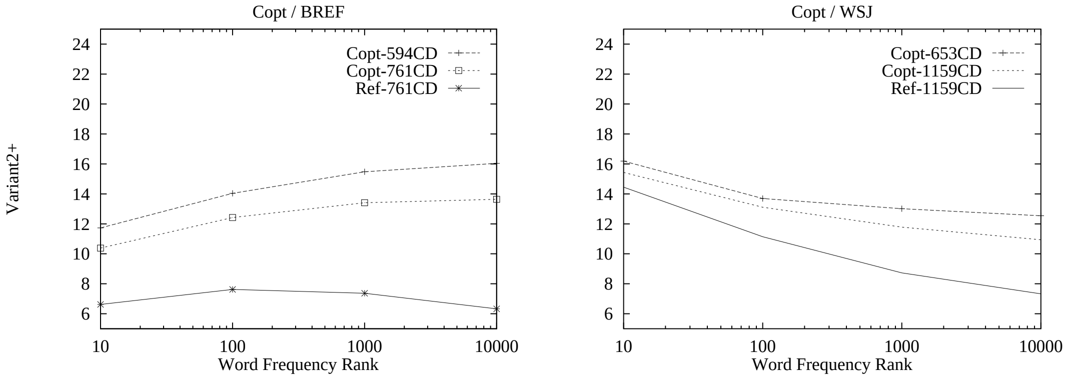

Figure 4: *Variant2+* rate vs Rank for French (BREF) and English (WSJ) using the *Copt* lexica and different acoustic models.

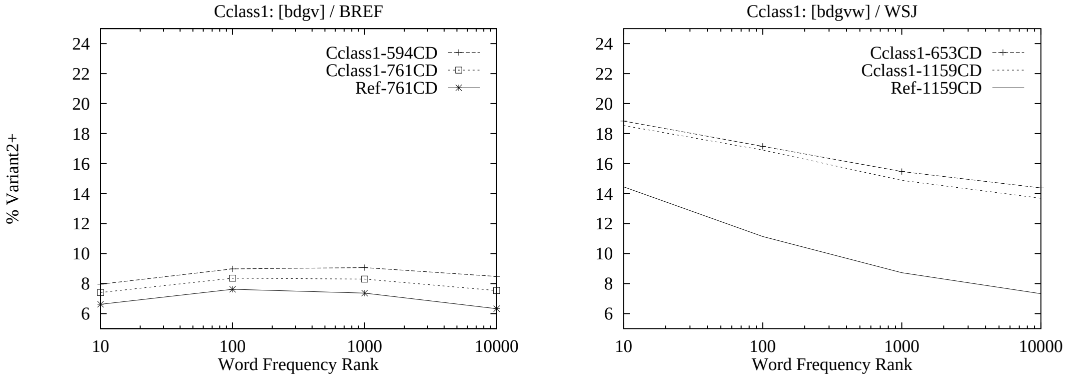

Figure 5: *Variant2+* rate vs Rank for French (BREF) and English (WSJ) using the *Cclass1* ([bdgv], [bdgvw]) lexica and different acoustic models.

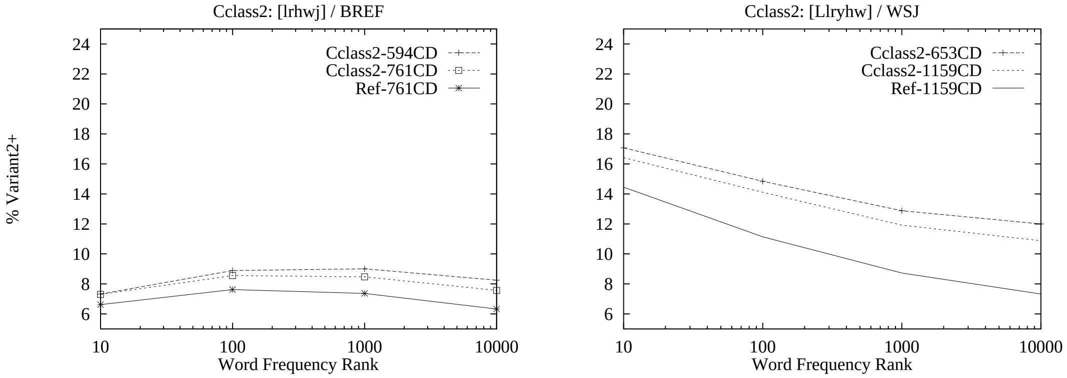

Figure 6: *Variant2+* rate versus frequency rank for French (BREF) and English (WSJ) using the *Cclass2* ([lrhwj],[Llryhw]) lexica and different acoustic model sets.

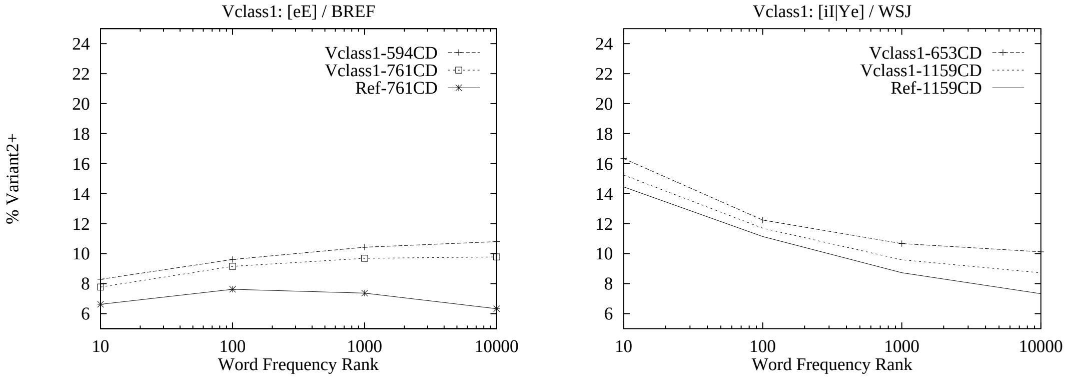

Figure 7: *Variant2+* rate versus frequency rank for French (BREF) and English (WSJ) using the *Vclass1* ([eE],[iI—Ye]) lexica and different acoustic model sets.

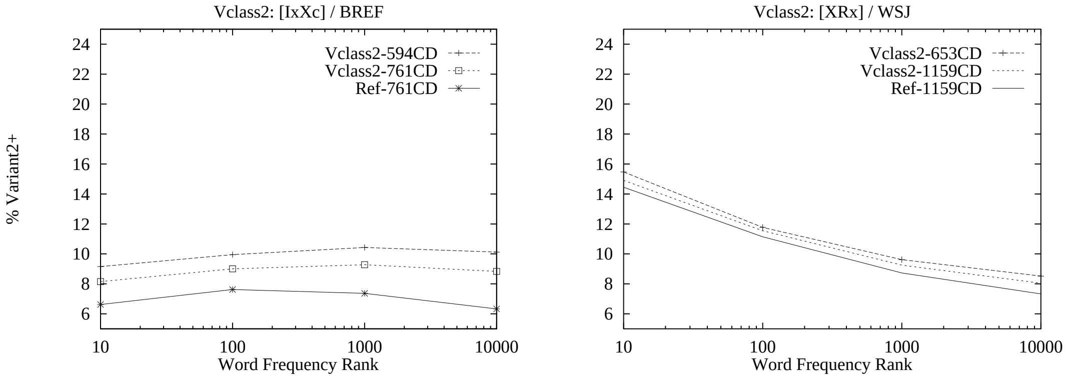

Figure 8: *Variant2+* rate versus frequency rank for French (BREF) and English (WSJ) using the *Vclass2* ([IxXc],[XRx]) lexica and different acoustic model sets.

10 most frequent words, a smaller variant rate of 1.3% is obtained for the BREF reference lexicon, compared to 2% for the MASK reference lexicon. This may explain the different behavior of the acoustic models sets trained. For the frequent words, the BREF models appear to be more word-specific, and the MASK and WSJ models to be more phone-specific. Given the limited vocabulary size of MASK, CD models tend to become rapidly word-specific.

Comparing MASK and BREF (see Fig. 9), the *variant2+* rates are found to be much higher for spontaneous speech when using CI acoustic models. The use of CD models tends to smooth the difference between the two different speaking styles.

We examined the subset of words ending in a Plosive-Liquid consonant clusters in BREF (25k words) and MASK (7k words), so as to measure the importance of the variant2+ rate in a context where a high percentage of sequential variants are expected. For read speech using *Copt* lexica and CI models, 38% of the words in this subset of BREF have been aligned with rank 2 and higher variants, compared to 51% for spontaneous speech. Concerning the occurrence of the word-final schwa in this context, it is much more frequent in read speech (65%) than in spontaneous speech (20%).

DISCUSSION AND PERSPECTIVES

The alignment results obtained with the above lexica have been used to study the link between word frequencies and variants using different acoustic model sets. We distinguished between the sequential and parallel variant types to investigate temporal and spectral modeling problems. We have introduced the *variant2+* rate to measure the representativity of the acoustic word models. We consider a decreasing *variant2+* rate with word frequency rank to be desirable for both linguistic reasons and from the point of view of lexical design for speech recognition: as infrequent words are not favored by the language model, they need accurate acoustic models in order to be identified.

The presented work can be considered as framework for more detailed linguistic analyses. Another aspect of future work aims at taking into account the presented observations in lexicon and acoustic modeling development and measure their impact in recognition experiments.

REFERENCES

[1] L.Lamel, G.Adda, “On Designing Pronunciation Lexicons for Large Vocabulary, Continuous Speech Recognition”, *ICSLP’96.*

[2] L.F. Lamel, J.L. Gauvain, M. Eskénazi, “BREF, a Large Vo-

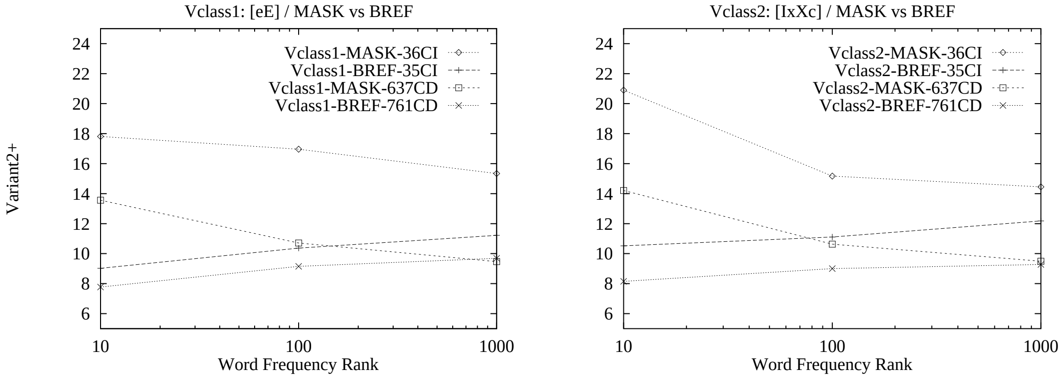

Figure 9: Comparison of the *Variant2+* rate versus word frequency rank on read (BREF) and spontaneous (MASK) speech in French using the *Vclass1* (left) and *Vclass2* (right) lexica.

cabulary Spoken Corpus for French," *EuroSpeech'91*.

[3] L. Lamel et al., "Development of Spoken Language Corpora for Travel Information", *EuroSpeech'95*.

[4] D.B. Paul, J.M. Baker, "The Design for the Wall Street Journal-based CSR Corpus," *ICSLP'92*.