Pronunciation Variants Across System Configuration, Language 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`

Number of pages: 24

Keywords: automatic speech recognition, lexical modeling, pronunciation variants, acoustic modeling

Published in *Speech Communication*, vol 29(2-4):83-98, Nov 1999.

Contents

List of Tables

List of Figures

ABSTRACT

This contribution aims at evaluating the use of pronunciation variants for different recognition system configurations, languages and speaking styles. This study is limited to the use of variants during speech alignment, given an orthographic transcription of the utterance and a phonemically represented lexicon, and is thus focused on the modeling capabilities of the acoustic word models. To measure the need for variants we have defined the *variant2+* rate which is the percentage of words in the corpus *not* aligned with the most common phonemic transcription. This measure may be indicative of the possible need for pronunciation variants in the recognition system.

Pronunciation lexica have been automatically created so as to include a large number of variants (overgeneration). In particular, lexica with parallel and sequential variants were automatically generated in order to assess the spectral and temporal modeling accuracy. We first investigated the dependence of the aligned variants on the recognizer configuration. Then 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 was made for French based on alignments of BREF (read) and Mask (spontaneous) data. Comparative alignment results using different acoustic model sets demonstrate the dependency between the acoustic model accuracy and the need for pronunciation variants. 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.

Cette contribution vise à évaluer l'utilisation des variantes de prononciation pour différentes configurations de système, différentes langues et différents types d'élocution. Cette étude se limite à l'utilisation de variantes pendant l'alignement automatique de la parole étant donnée une transcription orthographique correcte et un lexique de prononciation. Nous focalisons ainsi notre étude sur la capacité des modèles acoustique des mots à rendre compte du signal observé. Pour évaluer le besoin de variantes nous avons défini le taux de *variant2+* qui correspond au pourcentage de mots du corpus qui ne sont pas alignés avec la meilleure transcription phonémique. Ce taux peut être considéré comme indicatif d'un éventuel besoin de variantes de prononciation dans le système de reconnaissance.

Différents lexiques de prononciation ont été créés automatiquement générant différents types et quantités de variantes (avec surgénération). En particulier des lexiques avec des variantes parallèles et séquentielles ont été distingués afin d'évaluer la précision de la modélisation spectrale et temporelle.

Dans une première étape nous avons montré le lien entre le besoin de variantes de prononciation et la qualité des modèles acoustiques. Nous avons ensuite comparé différents phénomènes de variantes pour l'anglais et le français sur des grands corpus de parole lue (WSJ et BREF). Une comparaison entre parole spontanée et parole lue est présentée. Cette étude montre que le besoin de variantes diminue avec la précision des modèles acoustiques. Pour le français, elle permet de révéler l'importance des variantes séquentielles, en particulier du e-muet.

1 Introduction

Pronunciation variants can be related to a variety of factors such as the speaking style, speaking rate, individual speaker habits and dialectal region. Adding pronunciation variants in a recognition 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, the 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 are also potential error sources. We are therefore very careful when introducing variants in our lexicons used for automatic speech recognition (Lamel and Adda, 1996).

Capturing pronunciation variants has attracted researchers for many years. Some early work was concerned with determining and using phonological rules (Cohen and Mercer, 1974; Oshika et al., 1975; Shoup 1980). With the availability of large spoken corpora there has been renewed interest in pronunciation and phonological modeling, with particular interest in automatic determination of pronunciation variants (see for example, Cohen, 1989; Lamel and Gauvain, 1992; Riley and Ljojle, 1996; Jelinek, 1996), and studying and modeling variations in speaking rate (Mirghafori, Fosler and Morgan, 1995; Fosler et al., 1996).

Speech recognition systems can be used to obtain data with which a linguistic analysis of pronunciation variants in large speech corpora can be carried out. When speech data are aligned with acoustic word models which allow for pronunciation variation, the observed alignments provide frequencies for the main variants in the corpus (as relevant for the acoustic modeling component of the speech recognition system). The alignment results do of course depend on the acoustic models, and more generally on the parameters of the speech recognizer (such as phone or silence penalties). We can then try to explain the observed variants on a linguistic level, by the characteristics of the speech data, or on a speech engineering level, by the properties of the acoustic models.

In this contribution we examine the use of pronunciation variants during speech alignment, focusing on the appropriateness of the acoustic word models given the observed acoustic data. A first study, on the speech engineering level, aims at measuring the impact of the precision of the acoustic models on the use of pronunciation variants during alignment. This is done by comparing the alignments obtained using context-independent (CI) and context-dependent (CD) acoustic model sets. The use of automatically aligned pronunciation variants is then investigated along two different linguistic axes: the language (French and American English) and the speaking style (read or spontaneous).

Different types of pronunciation variants were automatically generated and included in the pronunciation lexica used for alignment. Variants are distinguished as sequential or parallel. Sequential variants allow some phonemes to be optional, hence increasing temporal modeling flexibility. Parallel variants allow alternative phonemes from an a priori defined subset to replace a given phoneme. These enable us to study the discriminability of acoustically similar phone models. Some of these variants have clear linguistic motivation. For example, in French the insertion or deletion of the schwa-vowel (usually in word-final position) is a major phenomenon of sequential variation.

2 Pronunciation variants and speech recognition

For automatic speech recognition two somewhat antagonistic goals have to be considered concerning pronunciation variants. The first goal is to increase the accuracy of the acoustic models, and the second is to minimize the number of homophones in the lexicon. As a general rule, if pronunciation variants increase homophone rates, word error rates are likely to increase despite better acoustic modeling. It is nonetheless important that the lexicon contain multiple pronunciations for some of the entries. These are evidently needed for homographs (words spelled the same, but pronounced differently) which reflect different parts of speech (verb or noun) such as `excuse`, `record`, and `produce`. An alternative is to include part of speech tags in the lexicon to distinguish the different pronunciations for the same graphemic form. Alternate pronunciations should also be provided when there are either dialectal or commonly accepted variants. One common example is the suffix `-ization` which can be pronounced with a diphthong (/ɑ^{j}/) or a schwa (/ə/). Another example is the palatalization of the /k/ in a /u/ context resulting from the insertion of a /j/, such as in the word `coupon` (pronounced /kupɑn/ or /kjupɑn/) as shown in Figure 1. By explicitly taking into account these alternative types of pronunciations in the lexicon, the acoustic models will be more accurate.

*** Figure 1 here ****

Figure 2 shows two examples of the word `interest` by different speakers reading the same text prompt: `In reaction to the news, interest rates plunged....` The pronunciations are those chosen by the recognizer during segmentation using forced alignment. In the example on the left, the /t/ is deleted, and the /n/ is produced as a nasal flap. In the example on the right, the speaker said the word with 2 syllables, without the optional vowel and producing a /tr/ cluster. Segmenting the training data without pronunciation variants is illustrated in the upper aligned transcription. Whereas no /t/ is observed in the first example, two /t/ segments had to be aligned. The aligned transcription obtained using a pronunciation dictionary including all required variants is shown in the bottom. This better alignment will result in more accurate acoustic phone models.

*** Figure 2 here ****

Pronunciation variants allowing a change of the number of phonemes from the canonical pronunciation seem to be of particular importance, as a severe temporal mismatch between the observation and acoustic word model often results in a recognition error. In our work in large vocabulary continuous speech recognition in French, we have observed that many errors are due to missing *liaison* at word boundaries (Adda et al., 1997b).

The most common liaison in French is made by inserting the phoneme /z/ after words ending with an `-s` or an `-x` which precede a word starting with a vowel. Stated as such, this rule is too general. The liaison phenomenon should be applied only within phrases, and not across phrase boundaries. Liaison is more frequent between articles and nouns, than between nouns and adjectives. Liaison is rarely made with adverbs, but can be found on adverbs of quantity before adjectives, for example in the word sequence `plus ouvert`. While not prohibited, successive liaisons are generally avoided by speakers. Two example errors involving liaisons made by our French AUPELF'97 system (Adda et al., 1997a) are shown in Table 1. Both errors are due to a missing liaison phoneme /z/. The lower part of Table 1 gives the phonemic transcriptions in the recognition lexicon.

*** Table 1 here ****

These examples illustrate that missing phonemes in an acoustic word model (formed by concatentating phone models according to the pronunciation in the lexicon) may introduce errors. The system may choose a solution which respects the observed consonant-vowel structure of the data, by exchanging vowels (/e,i/ of `écrites` are replaced by /ə,ɛ/ respectively) or consonants (/z,g/ of `-s anglais` are replaced by /s,b/ of `semblait` respectively).

In earlier work, we experimented with the straightforward solution of adding optional liaison phonemes to all words. Unfortunately this exhaustive approach did not reduce the word error rate as the large number of variants introduced additional homophone sequences and entailed different errors. To give an idea of the magnitude of the problem, over 25% of the words of the French vocabulary used in this work could have a /z/ liaison.

*** Table 2 here ****

A similar problem in French arises with the optional word-final schwa. When a schwa is present in the acoustic observation and but is missing in the lexicon, an insertion of a small function word (article, conjunction) is often observed. Some example errors involving *word-final schwas* are shown in Table 2. The schwa can be observed even if the orthographic form of the word does not have a *word-final* `-e`. While this vocalic segment appears most often after final consonants, it is rather common in the Parisian dialect to observe a *word-final schwa* appended to phrase-final vowels.

Acoustic segments (often schwa or consonants in complex clusters) can be missing in the speech signal, particularly if the word or word sequence is easily predictable by higher level knowledge. For example, the word sequences composing numbers and dates obey a restrictive syntax, thus at the acoustic level important reduction phenomena can occur without loss of intelligibility. The word-internal cluster /ndr/ in `hundred`, which occurs often in such sequences, can be substantially reduced. In French compound word sequences, such as `cen`<u>`tre`</u>` d' information`, `orches`<u>`tre`</u>` de chambre` may be significantly reduced losing up to the final three phonemes (corresponding to the final syllable `-tre` before `de`). Such reductions allow for an increase in the information flow rate without needing to reduce segmental durations, just by reducing the number of syllables. While rare in read speech, these reduction phenomena are quite common in spontaneous speech, and very difficult to handle for speech recognizers.

3 Speech corpora

Three corpora were used for these experiments as shown in Table 3. 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 corpus (Lamel, Gauvain, Eskénazi, 1991) contains 66.5k sentences (about 100 hours of acoustic data) from 120 speakers reading extracts of articles from the *LeMonde* newspaper. Although considerably more data are available for American English, in this work we have used a portion of the WSJ0 data (Paul and Baker, 1992) 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 (Gauvain et al., 1997). In this study we used 38k sentences from 409 speakers (35 hours of acoustic data) from the Mask corpus.

*** Table 3 here ****

Figure 3 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. 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 and 1000 words are seen to cover essentially the entire corpus.[1]

[1] For the alignment experiments described here fragments (incomplete utterances of words) observed in the

*** Figure 3 here ****

4 Pronunciation lexica

Starting with our standard pronunciation lexica (reference lexica) we have designed augmented pronunciation lexica allowing either for parallel or sequential variation. The variant choices have been motivated partially linguistically and partially based on an analysis of typical system errors. Our goal is twofold: first, to increase our insight into the spectral and temporal modeling accuracy or weakness of the acoustic models; and second, to identify major pronunciation variants occurring in large speech corpora.

For practical reasons, an upper limit of 100 variants per lexical entry was imposed. A word-final optional schwa vowel was added for all lexical entries in the sequential variant lexica for French, entailing in the following higher complexities for French sequential variant lexica.

4.1 Reference lexica

Some example entries from our reference lexica used for training acoustic models are shown in Table 4. These lexica typically contain 10% to 20% pronunciation variants needed to describe alternate pronunciations observed for frequent words (E1 in Table 4), 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) on frequent words.

speech data are included as separate lexical items in the different vocabularies. This introduces some entries with short phonemic transcriptions. While the volume of acoustic data representing word fragments remains very low, the number of entries with short phonemic transcriptions (one or two phonemes) grows significantly for spontaneous (Mask) data, explaining the knee in the curve for this data in Figure 4.

*** Table 4 here ****

4.2 Sequential variant lexica

*** Table 5 here ****

Large sequential variant lexica were automatically derived from the reference lexica (after systematically adding an optional word-final schwa) by allowing either all vowels or all consonants to be optional *Vopt* and *Copt*. Example entries for these lexica are shown in Table 5. These lexica aim at measuring the temporal modeling capabilities of the acoustic models. If a high proportion of temporal variants is observed, a lack of accuracy in the acoustic models is likely to be responsible. If the variant rate is low, the observed variants are probably due to real pronunciation variants. The *Vopt* lexicon, in addition to allowing for the well-known optional word-final schwas in French, 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 in Table 4). These are generalized in the *Copt* lexica.

4.3 Parallel variant lexica

*** Table 6 here ****

Parallel variant lexica have been generated by defining a variety of broad phoneme classes and allowing each phoneme in a given class to be replaced by any member of the same class. For each broad phoneme class a specific lexicon was generated. Table 6 lists the phoneme classes reported on here. The vowel classes are linguistically motivated, containing vowels in the same broad class which have been observed to be confusable by speech recognizers. In French, many quasi-homophones are separated by the open-closed distinction on vowels (e.g.: `est` /ɛ/, `et` /e/, verbs ending in `-er` /e/, past participle endings `-é` /e/, past tense endings `-ai,ais,ait` /ɛ/). In fluent speech the open-closed distinction may disappear, word identification relying increasingly on higher level constraints (lexical, syntactic, pragmatic, ...). A second vowel class in French contains the phonemes likely to be substituted for ə vowel segments in different contexts. For English, the first vowel class groups lax and tense front vowels, and the second contains the retroflex vowels and schwa.

The consonant classes were designed to evaluate the discriminative abilities of the corresponding acoustic models. The first class of consonants (*Cclass1*) contains a set of voiced plosives and weak fricatives (and /w/ for English). The second class focuses on liquids and glides.

4.4 Complexity of the variant lexica

*** Figure 4 here ****

*** Figure 5 here ****

For the purposes of this paper we define the complexity of the lexica to be the unweighted ratio of the *total number of variants* and the *total number of entries*. The complexity of the variant lexica depends on their word length distribution. Figure 4 displays the distribution of lexical items as a function of word length (in number of phonemes). For each lexical entry, the canonical pronunciation (assumed to be that with the most phonemes) is used to compute the word length. The left figure shows curves corresponding to the *Reference* lexica using the canonical pronunciations. The distribution for newspaper texts are seen to be quite similar for French and English. The spontaneous speech corpus contains on average shorter words, which is partially due to the presence of word fragments. When using the longest pronunciation in the Copt lexica, (Figure 4 right) the French curves are shifted to the right due to the word-final optional schwa vowel. Figure 5 gives an example of the average number of transcriptions in the *Reference* and *Copt* lexica. Whereas this number is close to 1 for the *Reference* lexica [2], the limit of 100 variants is achieved for the different *Copt* lexica with word lengths greater than 11.

*** Table 7 here ****

The complexity of the reference, sequential and parallel variant lexica are shown in Table 7 for each corpus. The *Copt* lexica have the highest number of variants for all corpora. The *Vopt* lexicon contains about one third of the number of variants as are in the *Copt* lexicon for English, and about one half for French. The larger numbers for BREF (cf. Table 3) are due to the word-final optional schwa. Since only instances of selected phonemes 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).

5 Measure for pronunciation variants

In this section we introduce the measure used to assess the use of variants in the aligned pronunciations, and how this measure is presented as a function of frequency rank of the

[2] Words containing 15 or more phonemes have a larger number of variants. These lexical entries are compounds and acronyms with phonemic transcriptions allowing for optional silences at word and letter boundaries.

observed lexical items.

5.1 Variant2+ rate

*** Table 8 here ****

We have chosen to measure the usefulness of pronunciation variants for improving acoustic modeling by counting the number of word occurrences aligned with alternate pronunciations. In particular, we define the *variant2+* rate as the percentage of word occurrences aligned with variants of frequency rank 2 or higher. This measure may be indicative of the need for pronunciation variants in the recognition system or equivalently of the inappropriateness of a unique acoustic word model generated with the most frequently used phonemic transcription. Table 8 gives examples of the *variant2+* rate for the French words *les, responsable* and the English words *hundred, economy.* The examples chosen have high *variant2+* rates. Although the second pronunciation for *hundred* can be considered canonical, it is seen to not be the most common.

5.2 Variant2+ curves

For each lexical item its *variant2+* rate is computed as shown in Table 8. The lexical entries are then sorted by their frequency rank in the speech corpora. The *variant2+* curves show the running average *variant2+* rates as a function of word frequency rank (see for example, Figure 6). A decreasing curve indicates that the less frequent words have fewer pronunciation variants, whereas an increasing curve results from higher *variant2+* rates on infrequent words.

Variant2+ curves are given in the next section for the different variant lexica. The curve for the corresponding *Reference* lexicon is always included. The gap between the *Reference* curve and the *Variant* curve indicates the importance of the particular phenomenon, which may be linked either to speech engineering factors or to linguistically motivated variants.

6 Experimental results

In this section we give experimental results aligning the different sequential and parallel lexica for different system configurations and combinations of language and speaking style. The results are displayed as curves showing the *variant2+* rate as a function of the word frequency rank. Linguistic and information theoretic intuition may suggest that acoustic variability can be higher for more predictable words. A simple approximation to word predictability is the word frequency.

With the exception of an experimental condition for French, The speech corpora used for the alignment experiments correspond to the speech corpora used for training. We are aware that this experimental setup probably underestimates the *variant2+* rate as expected from independent alignment data. However in these experiments our motivation was to look at variants on **large** corpora, providing a large number of occurrences per lexical entry. One of the system configurations described below incorporates only a small amount (about 8%) of the speech data for the acoustic model training (i.e. 92% of the aligned data is unseen). This configuration allows to approach the *variant 2+* rate behaviour on unseen data.

6.1 System configuration

To investigate the dependence of the choice of variant on the system configuration, alignment experiments were carried out using different acoustic model sets (see Table 9). There are 36, 35, 46 context-independent (CI) models respectively for Mask, BREF and WSJ0 corpora and around 650 context-dependent (CD-1) models for each of the three corpora. A second set of context-dependent (CD-2) models were used for read speech. For French, these were estimated on the Bref80 corpus, a 5.6k sentences subset of BREF, corresponding to about 10 hours of speech. Thus, there is a severe reduction in the amount of training data for the CD-2 model set compared to the CD-1 model set (10 hours versus 100 hours). For English, the number of contexts in the CD-2 model set is almost double that of CD-1 for a constant volume of training data. The French CD-2 models allow us to align 110 hours of unseen data, whereas the English CD-2 models allow us to study the *variant 2+* rate as a function of the number of contexts modeled. The context-dependent models, which are position-independent and allow for cross-word modeling, include triphones, left and right diphones and monophones. For information we provide the coverage of the triphone models on the speech corpora. Given approximately the same number of context-dependent models for WSJ0, BREF and Mask, the triphone model coverage is about 25% for read speech and 55% for spontaneous speech. The high coverage for Mask is due to the limited number of distinct lexical items (see Table 3). 1159 CD models yield a triphone coverage of 40% on WSJ0.

*** Table 9 here ****

*** Table 10 here ****

In Table 10 a significant decrease in the *variant2+* rate is observed with context-dependent (CD-1) 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. Increasing the number of CD acoustic models tends to reduce the need for pronunciation variants. The impact of the CD-2 models on the *variant2+* rate can be seen in Figures 6 through 11. Using the 594 CD-2 models for French (acoustic training material is reduced to 8% of the CD-1 data), the *variant2+* curves have the same global shape as the curves obtained with the CD-1 models, with an relative increase in variants between 10 and 20% depending on the type of variant lexicon. For English CD-2 models (1159 contexts) the *variant2+* rate is reduced by about 10 to 15% relative to the CD-1 model set (653 contexts) depending on the condition.

6.2 Language

*** Figure 6 here ****

*** Figure 7 here ****

In this section the *variant2+* rate for the French and English read speech corpora are compared, using different acoustic model sets and variant lexica. In Figures 6 and 7 the *variant2+* curves of the *Vopt* and *Copt* lexica are shown.

As can be expected from a priori linguistic knowledge about sequential variation in French (e.g. optional word-final schwa) our measures show a higher *variant2+* rate for French sequential lexica than for English. The *variant2+* rate for the French *Vopt* lexicon is about 13% as compared to about 6% for the *Reference* lexicon. In the *Reference lexicon* about 2.5% of variants (40%relative) are to be attributed to the word-final schwa. In the French *Vopt* lexicon, 7.6% (50% relative) of the *variant2+* rate can be attributed to this phenomenon. Whereas the *Vopt* lexicon offered many other possibilities (recall that all vowels were optional), the linguistically motivated variant of word-final schwa is observed to be important in improving the modeling accuracy. *Copt* lexica introduce a significant part of variants for both French and English. Concerning English, *Copt* makes use of significantly more variants than *Vopt*.

We can notice that the curves behave differently for the two languages. For French, the *variant2+* rate increases with frequency rank, whereas the corresponding English curves decrease substantially with frequency rank. The substantial differences in the sizes of the training corpora used for French and English do not account for the observed differences, since using French models trained on less data than English (594 models, CD-2 trained on

BREF80) and on more data than English (761 models, CD-2 trained on BREF) give very similar curves. The characteristics of the curves for English satisfy our previously stated intuition about acoustic variability and word frequency (predictability): the acoustic models seem to accurately represent phonemes, resulting in a larger variant rate for frequent words. In contrast, for French, the acoustic models seem to well represent the more frequent words generating few variants. However, more variants are observed for infrequent words. A possible linguistic explanation is that word-final schwas appear more often on less frequent content words and generate a significant part of the variants for these words in French. From speech engineering perspective one can argue that the acoustic phone models seem to be more word-dependent (for frequent words) and less phoneme dependent. 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.

*** Figure 8 here ****

*** Figure 9 here ****

*** Figure 10 here ****

*** Figure 11 here ****

Acoustic models for consonants are relatively accurate for French and less discriminative for English (see Figures 8 and 9). The opposite is observed for the vowel classes in the two languages (compare Figures 10 and 11). 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.

6.3 Read versus spontaneous speech

In this section we compare *variant2+* rates on the Mask and BREF corpora. The Mask *variant2+* rate curves are seen to globally decrease, as did the WSJ ones, which reflects that most frequent words have higher acoustic variability. 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 reference lexica (weighted lexica complexity). Considering only the 10 most frequent words, a smallerlexicon complexity of 1.3 is obtained for the BREF Reference lexicon compared to 2 for the Mask Reference lexicon. This may partially explain the different behavior of the acoustic model sets.

Comparing Mask and BREF (see Figure 12), the *variant2+* rates are much higher for spontaneous speech when CI acoustic models are used. The use of CD models tends to smooth out the differences between the two different speaking styles. But we have to recall here that, even if the number of acoustic models is comparable for Mask and BREF, the triphone coverage is over 50% on the spontaneous speech corpus and only 25% for the read speech corpus.

On a more linguistic level, we examined the subset of words ending in a Plosive-Liquid consonant cluster in BREF (25k words) and Mask (7k words), so as to be able 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. The word-final schwa in this context is much more frequent in read speech (65%) than in spontaneous speech (20%). This observation may also contribute to explain the difference in the read and spontaneous speech curves.

*** Figure 12 here ****

7 Discussion and Perspectives

Comparative alignment results using different acoustic model sets have demonstrated the impact of the acoustic modeling accuracy on the need for pronunciation variants. As the number of context-dependent models is increased covering more triphone contexts there is a reduced need for pronunciation variants in the lexicon. This observation, which is seen by the reduction in the *variant2+ rate*, provides insight from the speech engineering viewpoint.

The lexica generated for the alignment experiments reported here suffer from severe overgeneration. This overgeneration has been introduced on purpose to assess the acoustic modeling accuracy without needing to carrying out the more expensive phone recognition experiments on very large corpora. These lexica have allowed us to focus attention on specific problems of either linguistic or speech engineering interest.

We have distinguished between sequential and parallel variant types in order to investigate temporal and spectral modeling problems. The alignment results obtained with the above lexica have been used to study the link between word frequencies and variants with different acoustic model sets.

We have introduced the *variant2+* rate to measure the representativity of the acoustic word models. We consider that a *variant2+* rate that decreases with word frequency rank is desirable for both linguistic reasons and from the point of view of lexical design for automatic speech recognition: since infrequent words are not favored by the language model, they need accurate acoustic models in order to be identified.

This work can be considered as framework for more detailed linguistic analyses. In addition to such analyses, an important aspect of future work will be directed at taking into account the presented observations in lexicon and acoustic modeling development, and measuring their impact in recognition experiments.

8 References

G. Adda, M. Adda-Decker, J.L. Gauvain, L. Lamel, (1997a) "Le système de dictée du LIMSI pour l'évaluation AUPELF'97", *1ères JST FRANCIL*, Avignon, April 1997.

G. Adda, M. Adda-Decker, J.L. Gauvain, L. Lamel (1997b). "Text normalization and speech recognition in French," *Proc. ESCA Eurospeech'97*, Rhodes, Greece, **5**, pp. 2711-2714, September 1997.

M. Cohen, *Phonological Structures for Speech Recognition*, PhD Thesis, U. Ca. Berkeley, 1989.

P.S. Cohen and R.L. Mercer, "The Phonological Component of an Automatic Speech Recognition System, in D.R. Reddy, ed., Speech Recognition: Invited papers presented at the 1974 IEEE Symposium. New York: Academic Press, 1975.

E. Fosler, M. Weintraub, S. Wegmann, Y.-H. Kao, S. Khudanpur, C. Galles, and M. Saraclar, "Automatic learning of word pronunciation from data," *Proc. ICSLP'96*, Philadelphia, PA, **Addendum**, pp. 28-29, Oct. 1996.

J.L. Gauvain, S. Bennacef, L. Devillers, L. Lamel, R. Rosset: "Spoken Language component of the MASK Kiosk" in K. Varghese, S. Pfleger(Eds.) "Human Comfort and security of information systems", Springer-Verlag, 1997.

F. Jelinek, "DoD Workshops on Conversational Speech Recognition at Johns Hopkins, *Proc. DARPA Speech Recognition Workshop*, Harriman, NY, pp. 148-153, Feb. 1996.

L.F. Lamel, G. Adda, 1996. On Designing Pronunciation Lexicons for Large Vocabulary, Continuous Speech Recognition. *Proc. ICSLP'96*, Philadelphia, PA, **1**, pp. 6-9, Oct. 1996.

L.F. Lamel and J.L. Gauvain (1992), "Continuous Speech Recognition at LIMSI," *Proc. Final review of the DARPA ANNT Speech Program*, September.

L.F. Lamel, J.L. Gauvain, M. Eskénazi, 1991. BREF, a Large Vocabulary Spoken Corpus for French. *EuroSpeech'91.*

N. Mirghafori, E. Fosler, N. Morgan, "Fast speakers in large vocabulary continuous speech recognition: analysis and antidotes, " *Proc. Eurospeech-95*, Madrid, **1**, pp. 491-194, September 1995.

B.T. Oshika, V.W. Zue, R.V. Weeks, H. Neu, and J. Aurbach (1975), "The Role of Phonological Rules in Speech Understanding Research," *IEEE Trans. Acoustics, Speech, Signal Processing*, vol. ASSP-23, pp. 104-112.

D.B. Paul, J.M. Baker, 1992. The Design for the Wall Street Journal-based CSR Corpus. *ICSLP'92.*

M.D. Riley and A. Ljojle, "Automatic Generation of Detailed Pronunciation Lexicons", in *Automatic Speech and Speaker Recognition*, Kluwer Academic Pubs, Ch. 12, pp. 285-301, 1996.

J. Shoup, "Phonological Apsects of Speech Recognition," Chapter 6 in *Trends in Speech Recognition*, W.A. Lea, ed., Englewood Cliffs, NJ: Prentice-Hall, pp. 125-165.

	Reference transcription	System hypothesis
A	les plaintes écrites	les plaintes secrètes
	les industriels anglais	les industriels semblait

	Orthographic form	Phonemic form
B	plaintes	plɛ̃t
	écrites	ekrit
	secrètes	səkrɛt
	industriels	ɛ̃dystrijɛl
	anglais	ɑ̃glɛ ɑ̃glɛz
	semblait	sɑ̃blɛ sɑ̃blɛt

Table 1: Examples of recognition errors due to missing *liaison* phonemes. The reference transcription and the system hypothesis are provided in part **A**, the corresponding lexical entries in the French *Reference* pronunciation lexicon are shown in part **B**.

A

Reference transcription	**System hypothesis**
Bangkok	Bangkok que
publique	public que

B

Orthographic form	**Phonemic form**
Bangkok	bɑ̃kɔk
publique	pyblik
public	pyblik
que	kə

Table 2: Examples of recognition errors due to a missing *word-final schwa* phoneme. The reference transcription and the system hypothesis are provided in part **A**, the corresponding lexical entries in the French *Reference* pronunciation lexicon are shown in part **B**.

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

Table 3: Language, speaking style, total and distinct number of words for each corpus.

république	repyblik	F1
les	le lez	F2
prendre	prɑ̃dr{ə} prɑ̃d	F3
dix	dis{ə} di diz	F4
FOR	fɔr fɝ	E1
THAT	ð[æ,ə]t	E2
INVESTMENTS	ɪnvɛs{t}mən{t}s	E3
STEPHEN	stivən stɛfən	E4

Table 4: Example lexical entries for French (F1-F4) and English (E1-E4) in the reference lexica illustrating parallel ([]: alternate phonemes) and sequential ({}: optional phonemes) variants.

	Vopt	*Copt*
les	l{e}{ə} l{e}z{ə}	{l}e{ə} {l}e{z}{ə}
république	r{e}p{y}bl{i}k{ə}	{r}e{p}y{b}{l}i{k}{ə}
FOR	f{ɔ}r f{ɝ}	{f}ɔ{r} {f}ɝ
STEPHEN	st{i}v{ə}n st{ɛ}f{ə}n	{s}{t}i{v}ə{n} {s}{t}ɛ{f}ə{n}

Table 5: Example lexical entries in the *Vopt* and *Copt* lexica illustrating the augmented sequential flexibility.

	French	*English*
Vclass1	ɛ e	i ɪ ɨ ɑ^j e
Vclass2	ɛ̃ ə œ ɔ	ɜ˞ ɜ ə
Cclass1	b d g v	b d g v w
Cclass2	l r ɥ w j	ḷ l r j h w

Table 6: Phoneme classes for the parallel variant lexica design.

	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 7: Complexity of lexica: unweighted ratios $\frac{\#variants}{\#entries}$ in Reference, sequential *Vopt* and *Copt* lexica, and parallel *Vclass* and *Cclass* lexica.

Lexical entry	*rank*	*#occurences*	*variant2+*	*phonemic*	*#align*
les	3	21362	24%	le	16262
				lez	5100
responsable	471	205	47%	rɛspɔ̃sablə	109
				rɛspɔ̃sab	71
				rɛspɔ̃sabl	25
hundred	35	612	37%	hʌndɝd	387
				hʌndrəd	120
				hʌnɝd	89
				hʌnrəd	16
economy	382	60	47%	ɛkɑnəmi	32
				ikɑnəmi	28

Table 8: Example lexical entries in the Reference pronunciation lexica, frequency rank and number of occurrences in the speech corpora, *variant2+* rate, and the different phonemic transcriptions with the number of aligned occurrences (#align).

	Training Corpus			
	French			*English*
	Spont.	*Read*		
	Mask	*BREF*		*WSJ0*
Data used	*all*	*all*	Bref80	*all*
# *hours*	35	120	10	21
CI	36	35	-	48
CD-1	637	761	-	653
CD-2	-	-	594	1159

Table 9: Different acoustic model sets (context-independent CI and context-dependent CD) used for alignment. For each model set the amount of training data used is specified in number of hours. 'all' means that all the data used for the alignment experiments have been used for training (for most frequent phone units a randomly selected subset of limited size is used for training).

	French			*English*
	Spont.	*Read*		
	Model type	MASK	*BREF*	*WSJ*
Vopt	*CI*	22.2	18.6	15.7
	CD-1	13.0	12.9	9.9
Copt	*CI*	27.0	21.0	21.5
	CD-1	14.5	13.6	12.5

Table 10: Percentage of word occurrences aligned with a phonemic variant of frequency rank 2 or more (*variant2+* rate) for context-independent (CI) and context-dependent (CD-1) acoustic model sets using *Vopt* and *Copt* lexica.

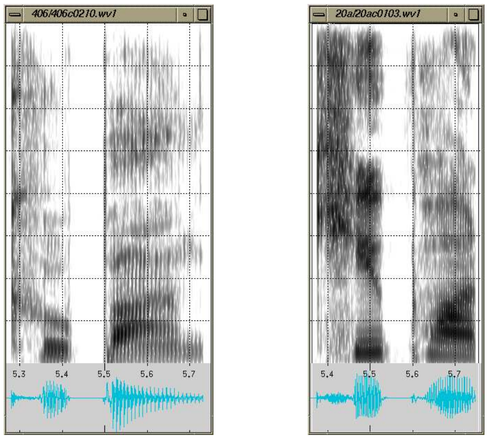

Figure 1: Spectrograms of `coupon`: /kupan/ (left, 406c0210) and /kjupan/ (right, 20ac0103).

The grid is 100ms by 1 kHz.

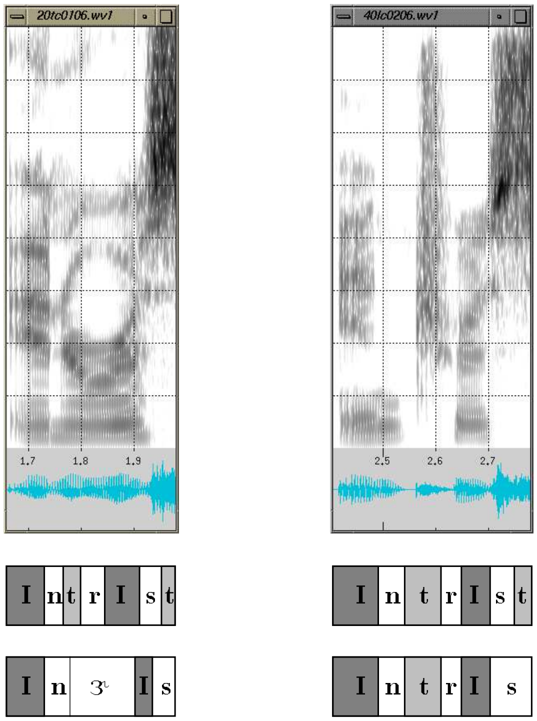

Figure 2: Spectrograms of the word `interest` with pronunciation variants: /ɪnɝɪs/ (left) and /ɪntrɪs/(right) taken from the WSJ corpus (sentences 20tc0106, 40lc0206). The grid is 100ms by 1 kHz. Segmentation of these utterances with a single pronunciation of `interest` /ɪntrɪst/ (upper transcription) and with multiple variants including /ɪntrɪs/ /ɪnɝɪs/ (lower transcription). The /ɪ/ and /t/ segments are light and dark grey respectively.

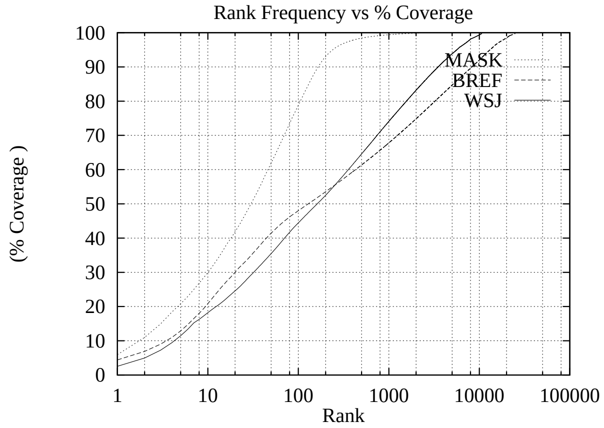

Figure 3: Lexical coverage of spontaneous speech (Mask) and read speech WSJ and BREF corpora as a function of word frequency rank.

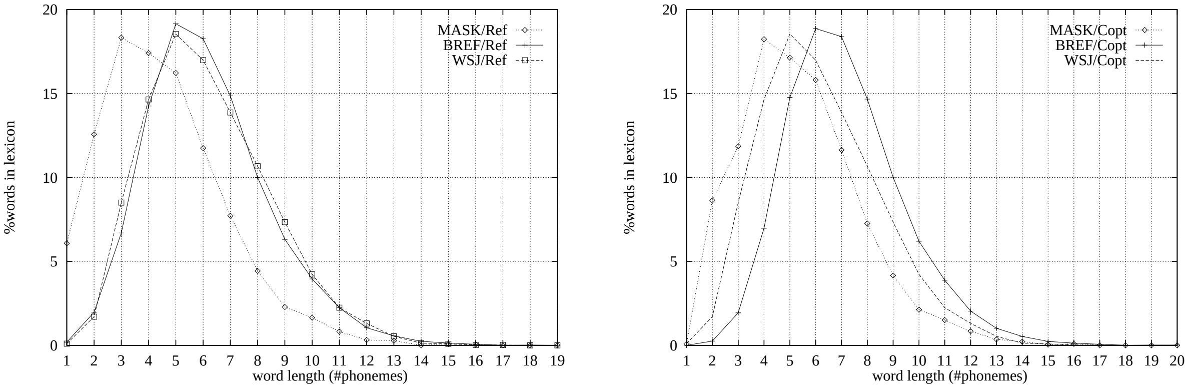

Figure 4: Distribution of lexical items as a function of length (in # of phonemes) of the canonical pronunciation in the *Reference* and *Copt* lexica.

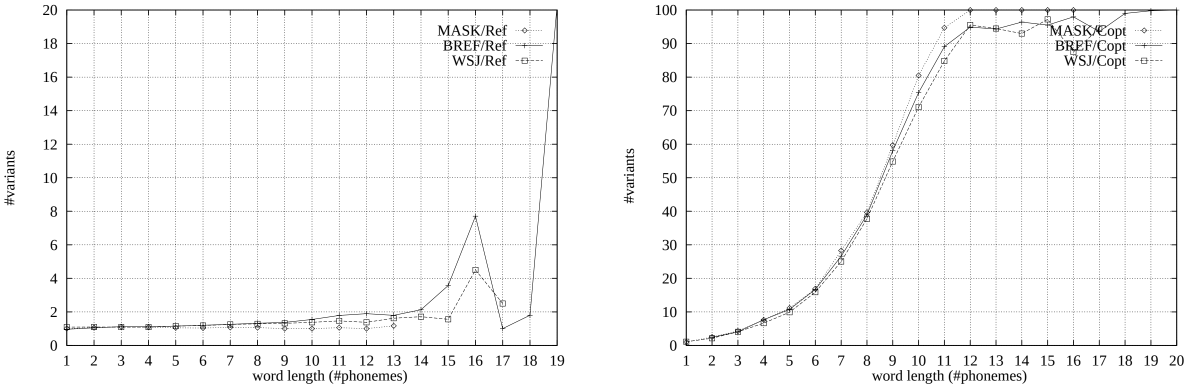

Figure 5: Average number of variants as a function of the length (in # of phonemes) of the canonical pronunciation for the *Reference* and *Copt* lexica.

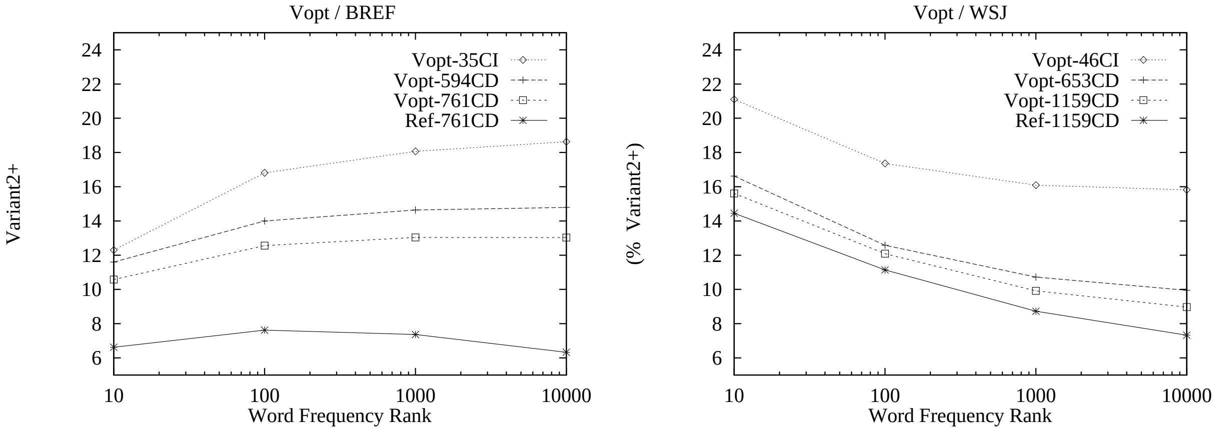

Figure 6: *Variant2+* rate vs Word Frequency Rank for French (BREF) and English (WSJ) using the *Vopt* lexica and different acoustic models.

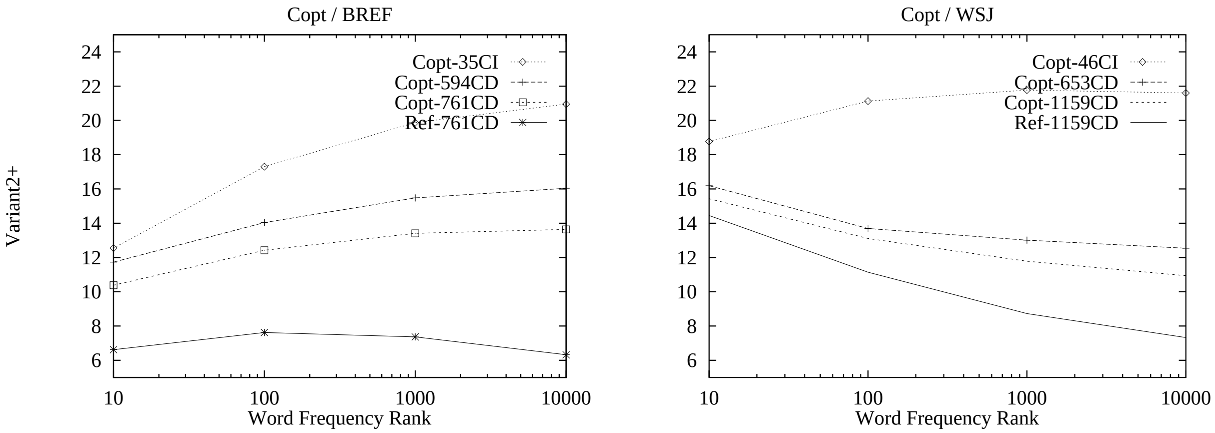

Figure 7: *Variant2+* rate vs Word Frequency Rank for French (BREF) and English (WSJ) using the *Copt* lexica and different acoustic models.

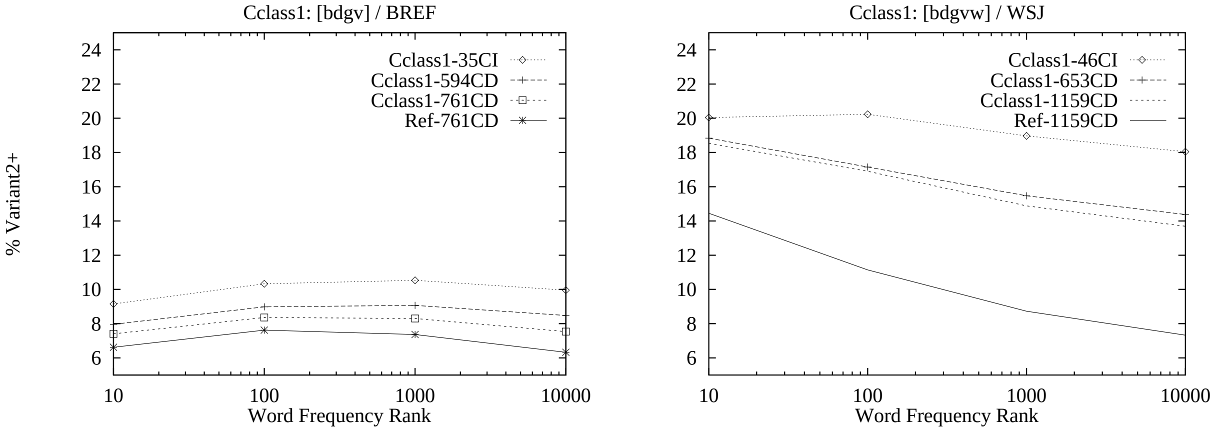

Figure 8: *Variant2+* rate vs Word Frequency Rank for French (BREF) and English (WSJ) using the *Cclass1* ([bdgv], [bdgvw]) lexica and different acoustic models.

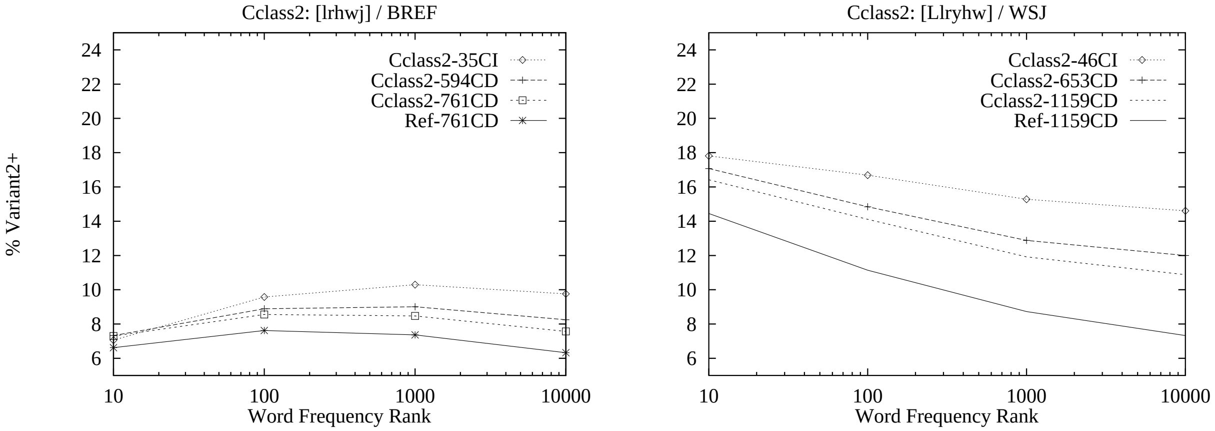

Figure 9: *Variant2+* rate versus frequency rank for French (BREF) and English (WSJ) using the *Cclass2* ([lrhwj],[ḷlryhw]) lexica and different acoustic models.

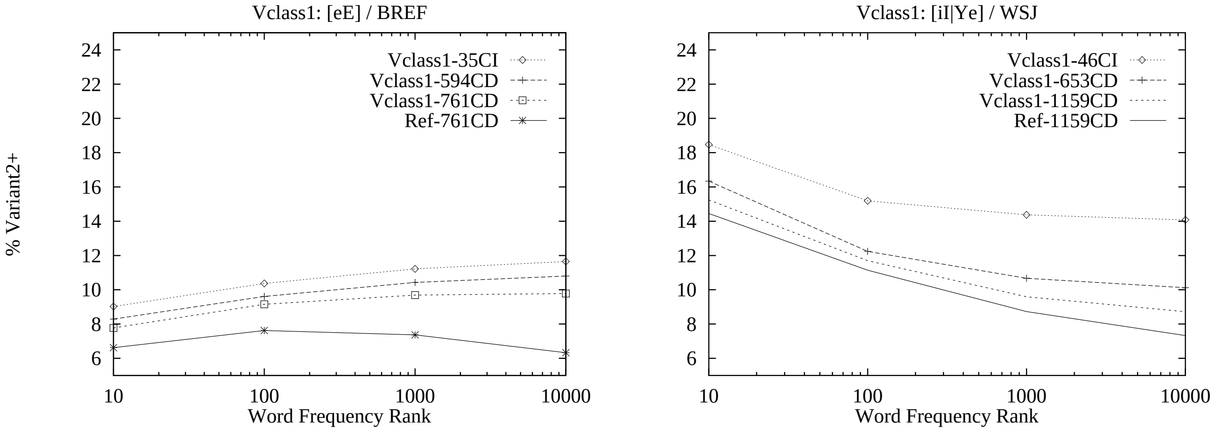

Figure 10: *Variant2+* rate versus Word Frequency Rank for French (BREF) and English (WSJ) using the *Vclass1* ([eɛ],[iIɨɑ^{j}e]) lexica and different acoustic model sets.

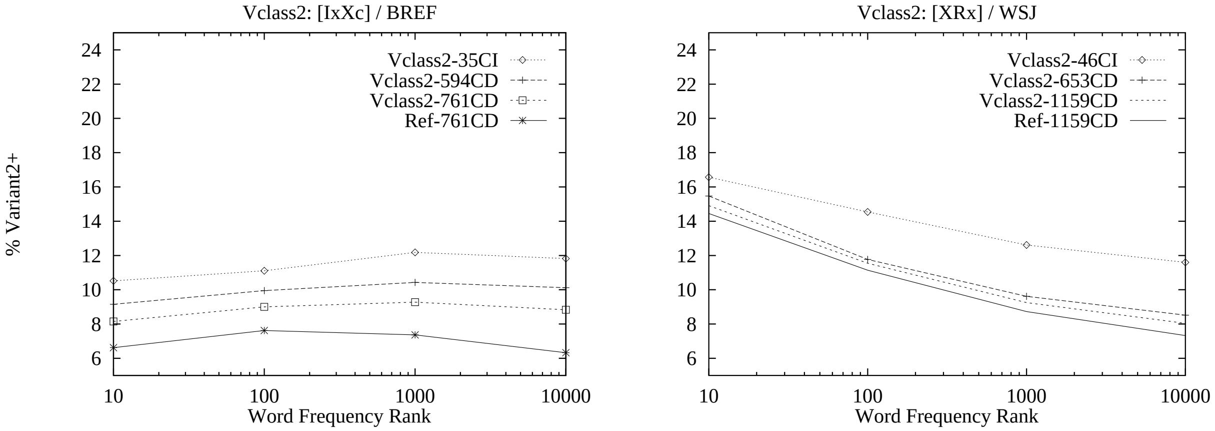

Figure 11: *Variant2+* rate versus Word Frequency Rank for French (BREF) and English (WSJ) using the *Vclass2* ([ɛ̃ ə œ ɔ],[ɝ ɜ ə]) lexica and different acoustic model sets.

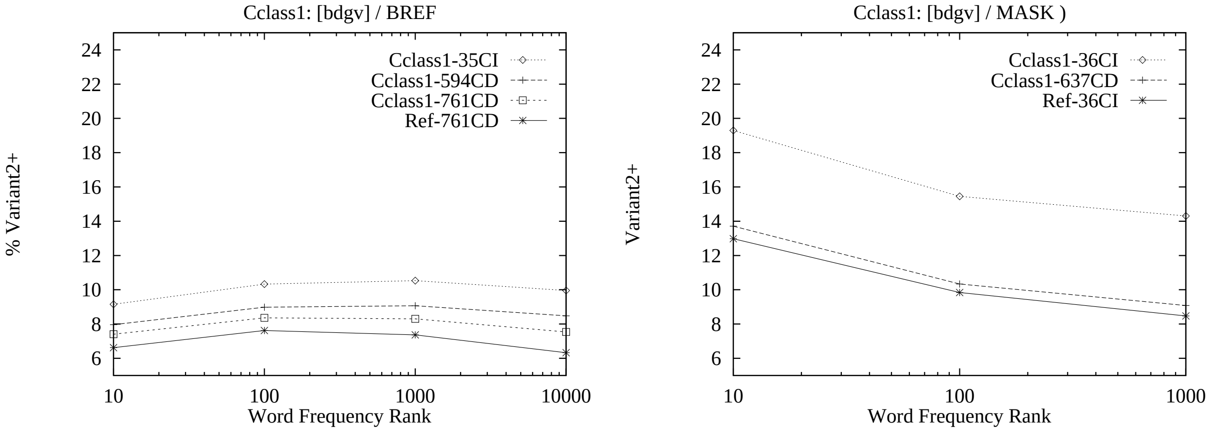

Figure 12: Comparison of the *variant2+* rate versus Word Frequency Rank on read (BREF) and spontaneous (Mask) speech in French using the *Cclass1* and *Vclass2* lexica.