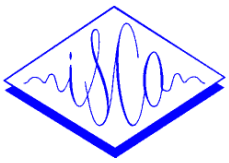

First International Conference on
Spoken Language Processing (ICSLP 90)
Kobe, Japan
November 18-22, 1990

Design Considerations and Text Selection for BREF, a large French read-speech corpus[1]

Jean-Luc Gauvain, Lori F. Lamel, and Maxine Eskénazi

LIMSI-CNRS
BP 133
91403 Orsay cedex, FRANCE

ABSTRACT

BREF, a large read-speech corpus in French has been designed with several aims: to provide enough speech data to develop dictation machines, to provide data for evaluation of continuous speech recognition systems (both speaker-dependent and speaker-independent), and to provide a corpus of continuous speech to study phonological variations. This paper presents some of the design considerations of BREF, focusing on the text analysis and the selection of text materials. The texts to be read were selected from 5 million words of the French newspaper, *Le Monde*. In total, 11,000 texts were selected, with an emphasis on maximizing the number of distinct triphones. Separate text materials were selected for training and test corpora. The goal is to obtain about 10,000 words (approximately 60-70 min.) of speech from each of 100 speakers, from different French dialects.

10.21437/ICSLP.1990-287

INTRODUCTION

One of the main obstacles to progress in continuous speech recognition has been the lack of sufficient speech material for the study of speech events and for training, development, and testing of algorithms and systems. A major effort in this area has been undertaken under the auspices of DARPA, with the production of the TIMIT[1,2] and Resource Management[3] speech corpora. The availability of these corpora has enabled speech recognition systems to be evaluated on a common ground, which has stimulated research in many laboratories, both within and outside of DARPA projects.

As a step in providing comparable data for the French language, we are recording BREF, a large read-speech corpus. The goal of BREF is to provide enough continuous-speech data for the development and evaluation of continuous speech recognition systems, particularly for the dictation task, and to provide a large enough corpus of read speech to be able to study and model phonological variations. Speech from at least 100 speakers will be recorded, so as to provide a broad basis for the study of speaker variability and data for development of speaker-independent recognition systems. With on the order of 10,000 words per speaker, it is hoped to provide enough speech data to study and model speaker-dependent characteristics.

In this paper the design considerations of BREF are described, focusing on the analysis and selection of text material, and the provision for separate corpora for training and testing. 11,000 texts to be read were selected from a source text of three months of the French newspaper *Le Monde*. The texts were selected using several criteria, including the requirement that they be fairly easy to read aloud. Other selection criteria tried to maximize the number of phonemic contexts and the number of different words. In addition, a set of sentences were selected each containing all the phonemes in French. Each speaker reads two of these sentences, which can be used for fast training and/or speaker adaptation.

[1]This work has been supported by the GRECO-PRC CNRS Communication Man-Machine program, ACCT (Agence de Coopération Culturelle et Technique), and ESPRIT POLYGLOT project.

TEXT SOURCE AND PREPROCESSING

The source text consisted of three months of *Le Monde* obtained from *L'Européenne de Données*. The text had been preformatted with codes to demarcate the title, author(s), and optional subject classification and credits. Thus, the first step in manipulating the text was to re-format it so as to eliminate the unnecessary information. The next step was to clean up the text by eliminating all incomplete sentences (this decision was based on punctuation) and correcting some text formatting errors. After these clean-up operations, approximately 4.2 million words of text remained. The "lost" text was roughly 50% header information and 50% textual errors. The final step in the preliminary processing used the punctuation markers to split the text into individual sentences, keeping the article and paragraph delimiters. Links to the original text were kept so that the source context of all sentences could be retrieved.

Also available were approximately 1.2 million words of Senate transcriptions which were used for comparison.

TEXT ANALYSIS

Each sentence was phoneticized using grapheme-to-phoneme rules[4], and erroneous pronunciations were hand-located[2] and corrected using an exceptions dictionary. The most common mispronunciations were foreign words and names, and acronyms. Also, each punctuation mark was replaced by a silence "phone."

Distributional properties

The distributional properties of the text were determined by counting the occurrences of sentence, word, and subword units. At the sentence level, counts were made of sentence types and

[2]Since this is such a labor-intensive procedure, corrections were made only for words occurring more than 20 times in the text.

lengths. At the word level, the number of distinct words and their word frequencies were counted. Subword units counted included syllables, dissyllables, phones, diphones, and triphones.

Sentence types: Sentences were classified as simple declarative, interrogative and exclamative types, or as more complex formulations which included ellipses, parenthetic expressions, and/or quotations. Table 1 shows the distribution of sentences in *Le Monde* according to type, and shows for each type the minimum, average, and maximum sentence lengths.

		Number of Words		
Sentence Type	*Percent*	*Ave*	*Min*	*Max*
Declarative	95	23	1	222
Interrogative	3.8	15	1	191
Exclamatory	1.2	13	1	104
Simple Sentences	57	19	1	191
Complex Sentences	43	33	3	222
Numbers	22	30	1	165
Acronyms	11			
Quotations	22	34	2	>400
Split Quotations	27	26	2	213
Parenthetic	11	35	2	>100

Table 1: Sentence types and lengths.

A conceptual problem was found while counting sentence types, a priori, a simple task: what should be done with end-of-sentence punctuation marks found within parenthetics or quotations? The analysis was performed two ways, ignoring and counting these marks. However, in sentence selection, it was decided to ignore parenthetic expressions as they are often too disjoint from the text, and to divide sentences within a long quotation into single, quoted sentences. This decision was made because sentences containing complex quotations could be quite long - over 500 sentences were found having more than 100 words each! While 12% of the quotations were only a single word and another 25% were 2-5 words long, the average length for a single quotation was 11 words. In contrast, parenthetics were typically short: over 75% had fewer than 5 words and the average length was 4 words.

Word and subword units: Word and subword units were counted in the phonemicized, syllabified text. Table 2 summarizes the counts for the different units for the complete text of *Le Monde* and the Senat. Counts made on a small subset of *Le Monde*, roughly 10%, showed the distributional properties of the text to be almost identical.

In the 167,359 sentences, there were almost 4.2 million words, over 90,000 orthographically distinct. To find the number of phonemic words, the grapheme-to-phoneme mapping was redone without the liaison rules, so as to avoid the ambiguity in word segmentation introduced by liaison. There were 64,000 phonemically distinct words, almost 30% less than the number of orthographically distinct words, giving a measure of the number of homophones in French. In order to know if the percent of homophones was dependent upon the vocabulary size, the percent homophones in 2000 and 10,000 most common words were determined, and also found to be roughly 30%. The dissyllable is defined from the midpoint of one vowel to the midpoint of the next vowel, and therefore contains all the intervening consonants. This unit has been successfully used for speech recognition and speech synthesis in French[5,6], in part because French vowels are acoustically relatively stable over time.

Unit	*Le Monde*	*Senat*
#sentences	167,359	64,613
#words (total)	4,244,810	1,137,928
#orthographically distinct	92,185	26,807
#phonemically distinct	63,981	
#syllables (total)	6,903,017	1,956,423
#distinct syllables	9,571	
#distinct dissyllables	37,636	
#phones (total)	16,416,738	4,737,578
#distinct phones	35	35
#distinct diphones	1,160	1,105
#distinct triphones	25,999	17,079

Table 2: Distributional properties of word and subword units.

On the average, there were 2.3 phones/syllable, 3.2 phones/dissyllable (including both vowels), and 3.7 phones/word. The most common phone was /r/, accounting for 8.0% and 7.9% of all phone occurrences in *Le Monde* and Senat, respectively. In counting phones, the vowels /*a*/ and /a/, as in the words "pâte" and "patte", and the nasal vowels /œ̃/ and /ɛ̃/, as in the words "brun" and "brin", were not differentiated since these are phonemic distinctions found in only some speakers of current French. Most of the possible diphones were found to exist (1160 out of 1225, taking into account the silence "phone"), as were 60% of the possible triphones. Some of these gaps are truly indicative of the French language, while others may be due to insufficient data or the grapheme-to-phoneme rules. However, the number of triphones may actually be elevated, relative to "traditional French", since there are so many foreign words (mostly names) in the text source.

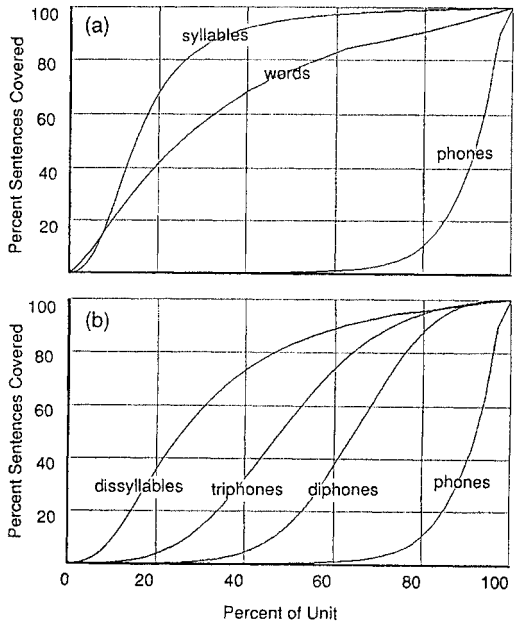

Figure 1: Frequency of occurrence for word and subword units.

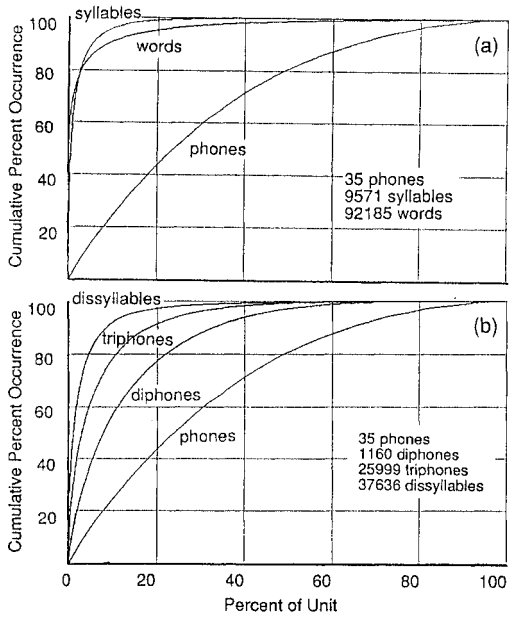

Figure 2: Percentage of sentences covered as a function of unit.

Figure 1 shows plots of the frequency of occurrence for the word and subword units in percentages. Part (a) has curves for words, syllables, and phones, and part (b) has curves for dissyllables, triphones, diphones, and phones. The units have been separated as such since words, syllables, and phones have no constraints internal to the unit itself restricting which units may follow, whereas the units in part (b) have internal constraints limiting the possible following units. Phones are shown in both for comparison as the basic unit.

Less than 20% of the distinct words account for over 95% of all word occurrences. In fact, 40% (about 35,000 words) occurred only once in the text, and 60% of the words appeared at most 3 times. This effect is even more pronounced for syllables, where the roughly 20% most common syllables account for 98% of all syllable occurrences. Almost 80% of the text is covered by only the most frequent 232 (20%) diphones. 20% of the triphones and dissyllables cover over 90% and 95% of the text, respectively.

But perhaps more interesting is the opposite question: given that 40% of the words only occurred once in the text, how many sentences can be pronounced if these words are eliminated? The curves shown in Figure 2 illustrate the percentage of sentences covered as a function of the percentage of word or subword unit. The curve for phones is very gradual - with 80% of the phones, only 10% of the sentences can be covered. For words, however, over 80% of the sentences are covered using only 60% of the distinct words, effectively eliminating all of the single occurrence words. The effect is even stronger for syllables: roughly 40% of the syllables cover over 90% of the sentences. Curves are shown for phones, diphones, triphones, and dissyllables in Figure 2b.

Entropy

In order to assess the relative importance of the word and subword units, the entropy of corresponding Markov sources were calculated. The probabilities used for each source are shown in Table 3a, where w_i, s_i, v_i, and a_i are respectively a word, syllable, vowel, and phone, and c_k is a string of consonants. A memoryless source was used to model the phone, word, and syllable sources. The diphone and dissyllable models were first order Markov sources, and the triphone model was a second order Markov source. All probabilities were estimated using frequency counts on the entire text.

(a) *Unit*	*order 0*	*order 1*	*order 2*
phonemic words	$p(w_i)$		
syllables	$p(s_i)$		
dissyllables	$p(v_i)$	$p(c_k, v_j \mid v_i)$	
phones	$p(a_i)$		
diphones	$p(a_i)$	$p(a_j \mid a_i)$	
triphones	$p(a_i)$	$p(a_j \mid a_i)$	$p(a_k \mid a_i, a_j)$

(b) *Unit*	*#distinct units*	*entropy (b/ph)*	*model I(b/ph)*
phonemic words	63,981	2.67	2.46
syllables	9,571	3.61	1.51
dissyllables	37,636	3.55	1.57
phones	35	4.72	0.40
diphones	1,160	3.92	1.21
triphones	25,999	3.40	1.72

Table 3: Markov sources: (a) model probabilities and (b) estimated entropies.

Table 3b summarizes the results of the models in bits/phone. The lowest entropies are found for the word and triphone sources, indicating that their models store the most information. Compared to the memoryless, equally probable 35 phone model, the information stored in the model is 2.46 and 1.72 b/ph, respectively.

TEXT SELECTION CRITERIA

Text selection was guided by the analysis performed and by text readability considerations. The texts were taken verbatim - none of the texts were modified, and each sentence was taken in its entirety. The first approach tried to choose sentences based on a measure of the information provided by the sentence. The idea was that sentences with relatively low information would be representative of the general text, and high information sentences would provide the less common events. Unfortunately, it was found that sentences with the most information were good predictors of foreign texts, and low information sentences all contained a date in the 1980's or 1990's.

Even more problematic, the size of the text source prohibited optimizing the criteria for sentence selection[7]. Thus, a more pragmatic approach was used. The text analysis indicated that words and triphones carried the most information. These two criteria gave approximately the same selection results, favoring one or the other by a small fraction. Using both constraints simultaneously was found to be too restrictive, limiting the total number of triphones. Since we believe triphones to be a more practical unit for training phone-based recognizers, we chose to maximize the number of triphones.

The readability of texts was assessed by asking people to read aloud selected sentences of a variety of types. The reading tests

included sentences of various lengths, from relatively short (10-20 words) to rather long (>60 words). The material included sentences with numbers, lists of numbers, acronyms, and quotations. In general, shorter sentences were easier to pronounce than longer ones. For sentences longer than 20 or so words, punctuation markers roughly every 10 words, greatly enhanced the sentence's readability. Sentences containing long lists of numbers or names also caused problems, as did acronyms. The problem presented by acronyms was that, if the acronym was unknown, people did not know whether to try to pronounce it as a word, or to read it as a list of letters.

Texts were selected in subsets, iteratively removing the already selected sentences. First, all of the sentences containing all of the phonemes in French were extracted, and 18 of these were hand-selected based on readability. Next 3 sets of each of the following types were selected. Paragraphs were chosen to provide semantic context for the speaker and to better model dictation task. The constraints on paragraph selection limited the number of sentences/paragraph (3-8), the number of words/sentence (5-24), and the number of new triphones/paragraph (>22). Short sentences (8-15 words) were selected since these are typically relatively easy to read, having a minimum of 8 new triphones and a maximum of 5 punctuation marks. Longer sentences (8-25 words) added a bit more diversity: two types were selected, the first requiring 12, and the second requiring 16 new triphones/sentence.

SELECTED TEXT SUBSETS

The selected texts consist of the 18 all phoneme sentences, and approximately 840 paragraphs, 3300 short sentences, and 3800 longer sentences. The paragraphs have an average of 4.2 sentences, each sentence having an average of 15.2 words. The average length of the short sentences is 12.4 words, and the longer sentences 21 words. The texts were separately selected in 3, roughly comparable sets: a set to be distributed for training, a second set to be distributed for test/evaluation purposes, and a third set to be kept undistributed (hidden) for a final, blind evaluation of systems. In these 3 sets, no sentence appears more than once; therefore, there is no overlap among the sets.

Unit	*Train*	*Test*	*Hidden*	*Total*
#sentences	3,877	3,624	3,501	11,002
#words (total)	55,760	50,946	49,040	115,746
#distinct words	14,089	12,803	12,280	20,055
#phonemic words	11,215	10,177	9,757	15,460
#syllables	3,339	3,040	2,976	3,977
#dissyllables	11,297	10,472	10,072	14,066
#phones (total)	252,369	230,102	222,250	726,988
#distinct phones	35	35	35	35
#diphones	1,107	1,092	1,082	1,115
#triphones	15,704	14,769	14,399	17,552

Table 4: Distributional properties of selected text subsets.

The distributional properties for the 3 sets of texts, and the combined total, are shown in Table 4. The sets are distributionally comparable in terms of their coverage of word and subword units and quite similar in their phone and diphone distributions. The most common phones have the same frequencies for all 3 sets: /r/ 7.8%, /a/ 7.2%, /l/ 6.0%, /s/ 5.8% and /i/ 5.4%. The most common diphones are /də/, /ar/, /la/, /ɛr/, and /tr/, and triphones are /sjɔ̃/, /par/, /dəl/, /trə/, and /asj/.

Subdivision of texts for individual speakers

Text subsets for individual speakers are automatically selected from the text sets. Each speaker reads 2 all phoneme sentences, 15 sets of 4 paragraphs, 10 sets of 18 short sentences, 10 sets of 14 long sentences, and 5 sets of 12 long, high density sentences. In total, each speaker reads an average of 650 sentences, comprised of about 9500 words. In these the speaker may be expected to cover roughly 3400 distinct orthographic words (3000 phonemic), 1400 syllables, 4300 dissyllables, 35 phones, 930 diphones, and 7500 triphones. In comparison to the distributional properties of a text set, each speaker pronounces a fair proportion of the subword units.

SUMMARY

Some of the design issues in choosing the contents of BREF have been presented, along with a summary of the text analysis. The selected materials maximized the number of distinct triphones. All the texts have been extracted verbatim from the original - no sentences were hand-designed or modified. Separate text materials with similar distributional properties were selected for training, testing, and hidden sets. Different speakers will record the materials providing non-overlapping speech corpora. Recording of BREF began in July 1990, and we expect to have at least 35 speakers recorded by November 1990.

ACKNOWLEDGEMENTS

We would like to thank Anne-Marie Derouault (IBM, France) for her efforts concerning BREF and the Senat for providing the texts of the Senat sessions.

REFERENCES

[1] W.M. Fisher, G.R. Doddington, and K.M. Goudie-Marshall, "The DARPA Speech Recognition Research Database: Specifications and Status," Proc. DARPA Speech Recog. Workshop, 1986.

[2] L.F. Lamel, R.H. Kassel, and S. Seneff,"Speech Database Development: Design and analysis of the acoustic-phonetic corpus," Proc. DARPA Speech Recog. Workshop, 1986.

[3] P. Price, W.M. Fisher, J. Bernstein, and D.S. Pallett, "The DARPA 1000-word Resource Managment Database for Continuous Speech Recognition," Proc. ICASSP, 1988.

[4] B. Prouts,"Contribution à la synthèse de la parole à partir du texte: Transcription graphème-phonème en temps réel sur microprocesseur", Thèse de docteur-ingénieur, Université Paris XI, Nov.1980.

[5] H. Singer and J.L. Gauvain, "Connected speech recognition using dissyllable segmentation," Fall meeting of the Acoust. Soc. of Japan,1988.

[6] J.L. Gauvain, "Le système de reconnaisance AMADEUS: Principe et algorithmes," LIMSI internal report, June 1990.

[7] A. Falaschi,"An Automated Procedure for Minimum Size Phonetically Balanced Phrases Selection," Proc. ESCA Workshop on Speech I/O Assessment and Speech Databases, 1989.