

BREF, a Large Vocabulary Spoken Corpus for French *

Lori F. Lamel, Jean-Luc Gauvain, and Maxine Eskénazi LIMSI-CNRS, BP 133
91403 Orsay cedex, FRANCE

ABSTRACT

This paper presents some of the design considerations of BREF, a large read-speech corpus for French. BREF was designed to provide continuous speech data for the development of dictation machines, for the evaluation of continuous speech recognition systems (both speaker-dependent and speaker-independent), and for the study of phonological variations. 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 selection criteria that emphasized maximizing the number of distinct triphones. Separate text materials were selected for training and test corpora. Ninety speakers have been recorded, each providing between 5,000 and 10,000 words (approximately 40-70 min.) of speech.

INTRODUCTION

One of the main obstacles to progress in continuous speech recognition has been the lack of sufficient speech material for the training, development, and testing of algorithms and systems, as well as for the study of speech phenomena. Several major efforts in this area have been undertaken under the auspices of DARPA, with the production of the TIMIT[1, 2] and Resource Management[3] speech corpora, and more recently the recording of ATIS[4]. 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 data will be recorded from at least 120 speakers (90 speakers have been recorded as of May 1991), so as to provide a broad basis for the study of speaker variability and data for development of speaker-independent recognition systems. In order to provide enough speech data to study and model speaker-dependent characteristics, each of 50 speakers have recorded approximately 10,000 words of text. The remaining speakers (40 of whom are already recorded) will each provide about half the speech data, reading about 5,000 words of text. We expect to have all the recordings completed by July 1991.

This paper describes some of the design considerations of BREF, including the recruitment of speakers, the recording environment, and the selection of the text materials. The speakers have been chosen from a pool of over 250 subjects in the Paris area. The recordings are made in stereo, in a sound-isolated room, and are monitored to assure the contents. The text materials were selected from a source text of three months of the French newspaper *Le Monde*. The selection criteria maximized the number of phonemic contexts and the number of different words, while requiring that the texts be fairly easy to read aloud. A set of 18 sentences were selected, each containing all the phonemes in French. Each speaker read two of these sentences which can be used for fast training or speaker adaptation. A preliminary analysis of some of these sentences is also given.

SPEAKER POPULATION

The speakers, chosen from a subject pool of over 250 persons in the Paris area, were paid for their participation. The subjects were given a short reading test, containing selected sentences from *Le Monde* representative of the type of material to be recorded. The experimenter evaluated the reading aptitude of the subject based on the following criteria:

- ability to follow instructions (ie. how to pronounce abbreviations, proper names ...)
- fluidity
- speaking rate
- number of errors (pronunciation, word substitution, hesitation, repetition)
- ease with task

and rated the subjects as “very good”, “good”, “poor”, and “very poor”. Table 1 shows the percentage of subjects in each category, based on the subject ratings. It was surprising that one-third of the subjects were judged to be “poor” or “very poor” at reading the texts. These subjects were not chosen for recording.

<i>Aptitude Rating</i>	<i>Percent of Subjects</i>
very good	40
good	28
poor	17
very poor	15

Table 1: Reading aptitude of subjects

Of the 90 speakers whom have been recorded, 50 are female and 40 are male. Table 2 gives the number of speakers recorded for each of the text subsets. 50 of the speakers recorded training material, and 20 development test material, and 20 test material to be used for evaluation purposes. Table 3 shows the number of speakers as a function of age and sex.

<i>Corpus</i>	<i>Number of Speakers</i>		
	<i>Male</i>	<i>Female</i>	<i>Total</i>
training	22	28	50
development	9	11	20
evaluation	9	11	20

Table 2: Speakers for training, development test, and evaluation test corpora

Over two-thirds of the speakers have a university education, with half having a specialized or advanced degree. Students account for 39% of the speakers, and professionals another 31%. All speakers were born in France except 4 speakers from Morocco and 2 speakers from Luxemburg. A database of subject information has been gathered, including information about the physical characteristics of the speaker (sex, age, date of birth, height, ...), the speaker’s background (parent’s languages, education, socio-economic level, ...), and the recording specifications (date, microphone, corpus, ...). The information in the speaker corpus will be distributed along with the speech corpus.

Age	Number of Speakers		
	Male	Female	Total
≤20	3	3	6
21-30	21	25	46
31-40	9	10	19
41-50	3	6	9
51-60	2	3	5
>60	2	3	5

Table 3: Age and sex of speakers

RECORDING ENVIRONMENT

The speech data has been recorded at LIMSI, using the recording facility *LIMREC* [5]. The recording setup consists of a “monitor” station and a talker station. The monitor has a PC-based workstation to control and monitor the recordings, headphones to listen to the recordings, and a microphone to communicate with the talker. The talker, located in an acoustically isolated room, reads the text prompts presented on a screen. The texts are presented in paragraph context when appropriate. Recordings are made in stereo using a close-talking, noise cancelling Shure SM10 and a table-top Crown PCC160 microphone. Also present in the room is a loud speaker through which the monitor may talk or play the recorded utterances for verification. When the monitor believes that the talker has made an error, the monitor requests that the talker listen to the utterance to verify it. If the talker detects the error, the text is re-recorded. If the talker determines that the recording is correct, the utterance is flagged, and the session is continued. Communication between the monitor and the talker is mediated by the facility, so as to prevent inadvertent recording errors. A log of the session is automatically generated containing a list of the filenames of the recordings and their durations, and the parameter values used during the session. The time to record a speaker is on the order of 6 to 8 hours, depending upon the speakers ease with the task, and the number of verifications/repetitions due to errors.

TEXT MATERIALS

Text source and preprocessing

The source text consisted of three months (approximately 5 million words) of *Le Monde* obtained from *L’Européenne de Données*. The text was reformatted to remove unnecessary header information, to eliminate incomplete sentences (based on punctuation), and to correct text formatting errors. Next, the punctuation markers were used 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. Each sentence was phoneticized using grapheme-to-phoneme rules[6], and erroneous pronunciations were hand-located¹ and corrected using an exception dictionary. The most common mispronunciations were foreign words and names, and acronyms. Also, each punctuation mark was replaced by a silence “phone.”

Distributional properties

Word and subword units were counted in the phonemicized, syllabified text. Subword units included syllables, disyllables, phones, diphones, and triphones. The disyllable 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[7, 8], in part because French vowels are acoustically relatively stable over time.

Counts for the different units are given in Table 4. Of the almost 4.2 million words of text over 90,000 are orthographically distinct.

¹Since this is such a labor-intensive procedure, corrections were made only for words occurring more than 20 times in the text.

However, there are only 64,000 phonemically distinct words – almost 30% less than the number of orthographically distinct words, reflecting the large number of homophones in French.²

Unit	<i>Le Monde</i>
#sentences	167,359
#words (total)	4,244,810
#orthographically distinct	92,185
#phonemically distinct	63,981
#syllables (total)	6,903,017
#distinct syllables	9,571
#distinct disyllables	37,636
#phones (total)	16,416,738
#distinct phones	35
#distinct diphones	1,160
#distinct triphones	25,999

Table 4: Distributional properties of word and subword units.

On the average, there are 2.3 phones/syllable, 3.2 phones/disyllable (including both vowels), and 3.7 phones/word. The most common phone is /t/, accounting for 8.0% of all phone occurrences. Most of the possible diphones exist (1160 out of 1225, taking into account the silence “phone”), as well as 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 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 disyllables, 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.

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 disyllables cover over 90% and 95% of the text, respectively.

Text selection criteria

The text materials were selected based on the text analysis and on readability considerations. The text analysis indicated that words and triphones carried the most information. (See [10] for more details on the text analysis and selection.) 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 and sentences containing numbers, acronyms, and quotations. Shorter sentences were easier to pronounce than longer ones. Punctuation markers are needed roughly every 10 words for sentences longer than 20 or so words, to be readable. Sentences containing long lists of numbers or names posed problems, as did acronyms. The problem presented by acronyms was that, if

²To 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.

average of 15.2 words. The average length of the short sentences is 12.4 words, and the longer sentences 21 words.

<i>Unit</i>	<i>Train</i>	<i>Dev.</i>	<i>Eval.</i>	<i>Total</i>
#all phoneme	18			18
#paragraphs	280	280	280	840
#sent in par	1433	1253	1187	3873
#short	1134	1080	1062	2276
#long(12)	812	812	784	2408
#long(16)	480	480	468	1428
Total	3877	3624	3501	11,002

Table 5: Number of sentences in text subsets.

<i>Unit</i>	<i>Train</i>	<i>Dev.</i>	<i>Eval.</i>	<i>Total</i>
#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 6: Distributional properties of selected text subsets.

The distributional properties for the 3 sets of texts, and the combined total, are shown in Table 6. 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 /sjø/, /par/, /dəl/, /rə/, 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.

DATA ANALYSIS

A preliminary data analysis has been carried out on the all phoneme sentences recorded by the first 20 training speakers. These sentences contain on average 41.2 words and 173 phonemes, and have an average duration of 15 seconds. The recordings have a signal-to-noise ratio of about 60 dB. Each sentence was listened to and the phonemic transcription was manually time-aligned. We observed in listening that the speaking style seemed to be a bit unnatural for some speakers - suggesting that these sentences were difficult to read aloud. Despite this, it was found that the speakers tended to follow standard pronunciation and punctuation protocols, ie. pauses are typically found at major syntactic boundaries. The main differences observed between the predicted pronunciation and what the speakers said stem from fast speaking rates, optional liaison, and optional pronunciation of the mute-e. The speaking rate ranged from 9.9 to 16.5 phonemes per second, with an average of 12.5.

These sentences were selected so as to provide, in a single utterance,

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

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. The texts were taken verbatim, with each sentence taken in its entirety. First, all of the sentences containing all of the phonemes in French were extracted, of which 18 were hand-selected based on readability. These sentences are meant to provide data for fast training and/or speaker adaptation. Next, paragraphs were selected so as to provide semantic context for the speaker and to better model the dictation task. Short sentences (8 to 15 words) were selected since these are typically relatively easy to read. Finally, longer sentences were selected to add diversity in the number of distinct triphones and words.

Selected text subsets

In total 11,002 texts were selected, including the 18 all phoneme sentences, 840 paragraphs containing 3872 sentences, 3276 short sentences, and 3836 longer sentences. The texts were separately selected in three, roughly comparable sets: a set to be distributed for training, a second set to be distributed for test/development purposes, and a third set to be kept undistributed for a final, blind evaluation of systems. There is no overlap of sentence texts in the three subsets. The speakers were also divided into training, development test, and evaluation sets, assuring no overlap in the speech material.

The number of sentences of each type is shown in Table 5. The paragraphs have an average of 4.2 sentences, each sentence having an

all the phonemes in French. The analysis has verified, that indeed, all the phonemes were present, and were pronounced.

SUMMARY

Some of the design considerations of BREF have been presented, including the speaker population, recording conditions, and text selection. The speakers' ages range from 20 to 65 years, and were recruited from the Paris area. Recordings are made in stereo, in an acoustically-isolated room using the recording facility *LIMREC*. Separate text materials with similar distributional properties were selected for training and testing. All the texts have been extracted verbatim from the original - no sentences were hand-designed or modified. Different speakers have recorded the materials providing non-overlapping speech corpora.

On a subset of sentences taken from the training speakers we verified that the recordings are correct and that speakers are able to provide read-speech data in this manner. Recording of BREF began in July 1990. At present 90 speakers have been recorded, and we expect to have recorded 120 speakers by July 1991.

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 Management Database for Continuous Speech Recognition," *Proc. ICASSP*, 1988.
- [4] C.T. Hemphill, J.J. Godfrey, G.R. Doddington "The ATIS Spoken Language Systems Pilot Corpus" *Proceedings DARPA Speech and Natural Language Workshop*, June 1990.
- [5] L.F. Lamel, "LIMREC Software: LIMSIS Recording Facility," *LIMSIS Internal Report*, 1990.
- [6] 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.
- [7] H. Singer and J.L. Gauvain, "Connected speech recognition using dissyllable segmentation," *Fall meeting of the Acoust. Soc. of Japan*, 1988.
- [8] J.L. Gauvain, "Le système de reconnaissance AMADEUS: Principe et algorithmes," *LIMSIS Internal Report*, June 1990.
- [9] A. Falaschi, "An Automated Procedure for Minimum Size Phonetically Balanced Phrases Selection," *Proc. ESCA Workshop on Speech I/O Assessment and Speech Databases*, 1989.
- [10] J.-L. Gauvain, L.F. Lamel, and M. Eskenazi, "Design Considerations and Text Selection for BREF, a large French read-speech corpus," *Proc. ICSLP*, 1990.