Proceedings of the International Conference on Acoustics, Speech, and Signal Processing, Detroit, Michigan, May 8–12, 1995.

PhoneBook: A Phonetically-Rich Isolated-Word Telephone-Speech Database

John F. Pitrelli, Cynthia Fong, Suk H. Wong, Judith R. Spitz, and Hong C. Leung

NYNEX Science & Technology, Inc., 500 Westchester Ave., White Plains, NY 10604 U. S. A.

ABSTRACT

We describe the collection of a phonetically-rich isolated-word telephone-speech database, "PhoneBook", which was undertaken because of (1) the lack of available large-vocabulary isolated-word data, (2) anticipated continued importance of isolated-word and keyword-spotting technology to speech-recognition-based applications over the telephone, and (3) findings that continuous-speech training data is inferior to isolated-word training for isolated-word recognition. PhoneBook has nearly 8000 distinct words, selected for complete coverage of phoneme contexts enumerated using both triphones and a novel method which takes into account syllable position, lexical stress, and non-adjacent-phoneme coarticulatory effects. PhoneBook consists of more than 92,000 utterances, averaging over 11 talkers for each word. A demographically-representative set of over 1300 native speakers of American English each made a single telephone call and read 75 words. This paper describes the word list design, talker enrollment procedure, recording procedure and equipment, utterance verification method, and summary statistics for PhoneBook, which will be made available through the Linguistic Data Consortium.

1. INTRODUCTION

Speech recognition research to date has focused primarily on wide-band speech. However, many anticipated uses for recognition are telephone-network applications, such as information services, financial transactions, and merchandise ordering. For the foreseeable future, the continued advancement of isolated-word and keyword-spotting telephone speech recognition capabilities will remain important, because (1) keyword recognition can aid many tasks and should be sufficient for portions of most applications, (2) it is possible to elicit isolated-word replies from a majority of users [1], and (3) isolated-word recognition will remain more robust and therefore important for handling noisy conditions even while continuous-speech systems emerge for more acoustically-favorable conditions.

As we have learned from experience with wide-band speech, large quantities of appropriate speech data are crucial to the development of speech recognition systems. Ideally, such databases would be created by having talkers dial in to a telephone-interfaced recording apparatus, preferably in the context of a realistic simulation of the anticipated speech-driven service. However, this is a time-consuming and expensive process, especially if it must be repeated for each new application. A short-cut is to play existing wide-band utterances across the telephone network, or process them through a filter which simulates the network, thus creating a "telephonized" copy. However, while phonetically-rich databases, such as TIMIT (wide-band) [2] and NTIMIT (telephonized) [3], exist for continuous speech, no isolated-word counterpart has been available. Current isolated-word databases tend to be limited to small application-oriented vocabularies, such as yes/no, digits, control words, etc. While one might consider using NTIMIT for large-vocabulary training, evidence shows that for isolated-word recognition, continuous-speech training data is inferior to using isolated words [4,5]. For development of large-vocabulary systems, a phonetically-rich isolated-word database analogous to TIMIT/NTIMIT is needed.

For these reasons, NYNEX has developed PhoneBook, a database consisting of over 92,000 utterances of almost 8000 words chosen to contain all phonemes in American English words in as many phonemic/stress contexts as are likely to produce coarticulatory variations. PhoneBook was also designed to span a representative variety of American dialects, talker characteristics, and telephone transmission characteristics, by employing a pool of over 1300 talkers chosen to be demographically representative of the United States.

2. DEVELOPMENT OF PHONETICALLY-RICH WORD LIST

The design of the phonetically-rich word list is central to the development of this database. The goal is to make the word list as compact as possible while (1) positioning each phoneme in a variety of phonetic and stress contexts wide enough to cover all significant coarticulatory variants, and (2) incorporating only words with single pronunciations which would be familiar to talkers from a wide variety of backgrounds. Our procedure is (1) filter words with any of a variety of pronunciation problems out of a large source dictionary, in order to obtain a list of "candidate" words for the database, (2) devise and apply to the candidate word list a criterion for enumerating phoneme sequences of length sufficient to capture coarticulatory effects, and (3) generate the final word list by choosing a subset of the candidate list in such a way as to capture each of the enumerated phoneme sequences at least once.

2.1. Source Dictionary

We began with two of the largest machine-readable dictionaries available, the 99,000-word CMU dictionary, and the 167,000-entry Moby Pronunciator. After eliminating multi-word entries, and reconciling phonemic representations into a common 42-phoneme inventory, shown in Table 1, we obtained a 146,742-word source dictionary. Vowels are marked with three levels of lexical stress.

i	bEAt	Y	bIte	n	Neat	D	THy
I	bIt	O	bOY	m	Meet	p	Pea
e	bAIt	W	bOUt	G	siNG	t	Tea
E	bEt	R	bIRd	h	Heat	k	Key
@	bAt	x	sofA	s	See	b	Bee
a	bOb	X	buttER	S	SHe	d	Day
c	bOUGHt	L	bottLE	f	Fee	g	Geese
o	bOAt	l	Let	T	THigh	C	CHurCH
^	bUt	w	Wet	z	Zoo	J	JuDGe
u	bOOt	r	Red	Z	meaSure		
U	bOOk	y	Yet	v	Van		

Table 1: Phoneme inventory, with example words.

This large dictionary's purpose was to be a source from which to draw words for a task in which obtaining a particular pronunciation from a naive speaker was of key importance. This purpose presumably diverges from the goals of dictionary makers; consequently, several categories of problematic words had to be filtered out:

1. Foreign words. This distinction is fuzzy, due to (1) varying levels of incompatibility between the word's foreign pronunciation and the pronunciation rules of English, (2) varying degrees of acceptance into English, and (3) varying degrees of Anglicization of the pronunciation during the assimilation process. The more assimilated, English-compatible and Anglicized words, such as "naive", were retained, while other words, such as "gendarme", were filtered out.
2. Difficult and obscure words. Some words would be pronounced inconsistently because they are unfamiliar and puzzling, such as "amanuensis" or "witenagemote". It is unlikely that we could include such a word for the purpose of obtaining any one particular pronunciation, and hope to get even a substantial fraction of talkers from the general public to provide that pronunciation.
3. Words with multiple acceptable pronunciations, whether the pronunciations are associated with different meanings, such as "produce", or not, such as "tomato". In either case, the word must be filtered out, as we cannot use such a word to elicit any one pronunciation reliably.
4. Potentially embarrassing words, such as those referring to private parts of the body. In addition to making talkers uncomfortable, they are unacceptable because they would likely disrupt some talkers' attention to the following words on their lists.
5. Words which are likely to be mis-read. Some words are much rarer than other similar-looking but differently-pronounced words. If we had permitted "carousal", a rare word stressed on the second syllable, to be on the list, many readers would likely have said "carousel", a common word stressed on the third syllable and therefore pronounced very differently.
6. Less common members of homonym pairs, such as "pidgin" and "Lett". These provide no advantage over the more familiar homonym.
7. Acronyms. We have chosen not to consider these to be words.

Proper names were retained, subject to the above criteria, as they effectively are English words, and are a good source of phoneme sequences. After eliminating problematic words, we obtained a filtered dictionary of 37,549 words to serve as a candidate word list.

2.2. Phoneme Sequence Enumeration and Word-List Development

Our criterion for phoneme sequence enumeration is based both on common practice in the speech recognition community and on acoustic-phonetic and articulatory knowledge. Most speech recognition research approximates phonemic contexts using triphones as equivalence classes, implicitly assuming that the primary coarticulatory effects on a phoneme are related to only the one preceding and one following phoneme. Some coarticulatory effects, however, reach beyond adjacent phonemes or are otherwise not covered by traditional triphone inventories. For example, the /u/ in "strewn" can cause some degree of anticipatory rounding throughout the /str/ sequence. Furthermore, lexical stress influences the acoustics of both vowels and consonants, but is typically not accounted for completely and consistently by triphone enumeration. Finally, the position within the syllable of a phoneme can affect the phoneme's articulation and acoustics.

For these reasons we enumerate phonemic contexts in terms of a simple syllable template, defining three "syllable parts" -- the "onset", consisting of all consonants preceding the vowel; the "nucleus", consisting of the vowel, a mark indicating one of three stress levels, and one postvocalic liquid if any; and the "coda", consisting of any remaining postvocalic consonants. However, to avoid reliance on the often-ill-defined syllable boundary within a sequence of consonants, we treat a coda-onset sequence as a single "part", with no loss of specificity in each enumerated context.

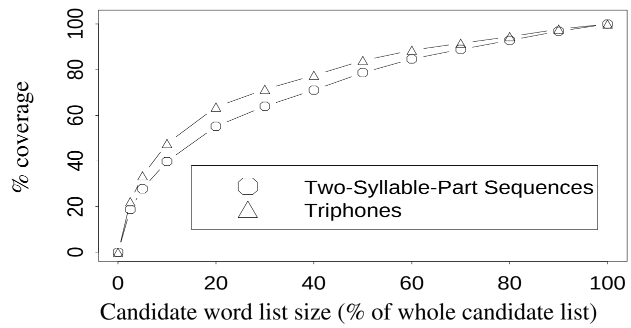

Figure 1: Coverage of contexts as a function of size of candidate word list. All values are expressed in terms of percent of values for full candidate word list (37,549 candidate words, which contain 9458 two-syllable-part sequences and 10,839 triphones).

Our inventory of phonemes-in-context contains all distinct sequences of two such syllable parts found in a dictionary. We also include a triphone inventory, due to the widespread interest in triphones in the recognition community. Beginning- and end-of-word were considered to be syllable parts and phonemes for purposes of developing the two-syllable-part-sequence and triphone inventories, respectively. Henceforth we refer to two-syllable-part sequences and triphones in aggregate as "contexts".

Working with the candidate word list, context enumeration yielded 10,839 triphones and 9458 two-syllable-part sequences. In order to verify that our source dictionary is of sufficient size to generate full coverage, we applied the context-enumeration algorithm to subsets of the candidate word list, in order to measure the degree to which the context inventories are approaching saturation. Figure 1 shows the number of contexts covered by fractions of the candidate word list, compared to the number in the whole candidate word list. As can be seen from the leveling of the curves, our source dictionary is large enough to encompass nearly all contexts that would have been found had we used a larger dictionary.

We then employed a greedy algorithm to extract a final word list as compact as possible from the candidate word list, while still covering our entire inventory of contexts. Our algorithm, which is similar to one described by Kassel [6], is as follows:

1. Identify each context that appears in only one word; accept those words into the final list and mark those contexts as covered.
2. Assign to each uncovered context a score which is the reciprocal of the number of candidate words which contain it.
3. Assign to each word a score which is the sum of the scores of its contexts.
4. Accept the highest-scoring word.
5. Set the scores of that word's contexts to zero; update scores of remaining candidate words.
6. Go to 4, unless all remaining candidate words score zero.

Word-list generation reduced the original 37,549-word candidate list down to a 7978-word final list while still containing every one of the contexts at least once. Because this algorithm favors compactness in terms of number of words, it tends to produce longer words than average text would contain. Our final word list aver-

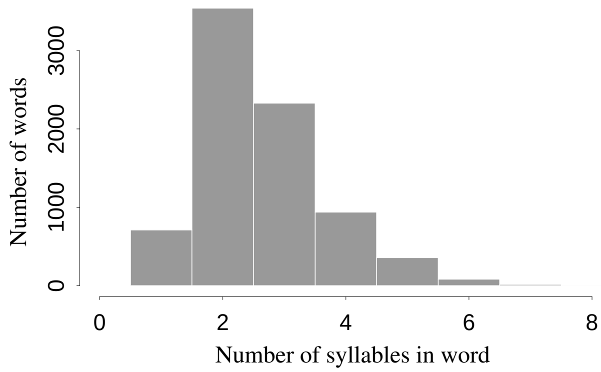

Figure 2: Histogram of number of syllables per word. Average is 2.62.

ages 2.62 syllables per word with a maximum of 8 (“counterrevolutionary” and “totalitarianism”); Figure 2 shows a histogram.

3. TALKER RECRUITMENT AND RECORDING CONDITIONS

A nationwide marketing-research firm selected a sample of prospective callers with the goal that response rates would provide a group of American talkers which is balanced for gender, and demographically representative of geographic location (within the 50 states and Washington, DC), income, age (over 18 years), education, and socio-economic status.

In order to make lists of a reasonable length for talkers to read, we sub-divided the final list into 106 lists of 75 or 76 words each. Several other items were added to the end of the lists, to make fullest use of 1300 new talkers; these include a read digit sequence, several prompts for spontaneous speech, and demographic questions. For purposes of this paper, these utterances are not considered to be part of PhoneBook.

The agency mailed the word lists to the callers, and entered talkers who completed the call into a lottery for savings bonds, in order to motivate participation. With this arrangement, 37% of the mailings resulted in completed calls.

A PC-based recording platform was developed in our lab for this and other similar data collections. The platform terminates a T1 trunk, enabling collection on up to 24 channels simultaneously. All data were digitally recorded by the platform in the T1 line’s 8-bit mu-law format, with no analog conversion.

Four channels of a T1 span were used for this data collection. Talkers were given a toll-free number connected to a hunt group for these channels. Talkers used their own telephone handsets to call our system.

Each list had three or four filler words added to the beginning to allow the speaker to become familiar with the recording procedure without consequence to the database. The caller was prompted for the word by the item number followed by a short beep, and was given three seconds to say the word. From the beginning of the greeting to the sign-off, each call lasted eight minutes and yielded 90 utterances, 75 or 76 of which are part of PhoneBook.

4. VERIFICATION

Once recorded, the utterances were checked for pronunciation accuracy, in order to verify that we had captured the intended phoneme sequences. For this stage of the collection, three transcribers were given phonetic training. Each transcriber was asked to verify

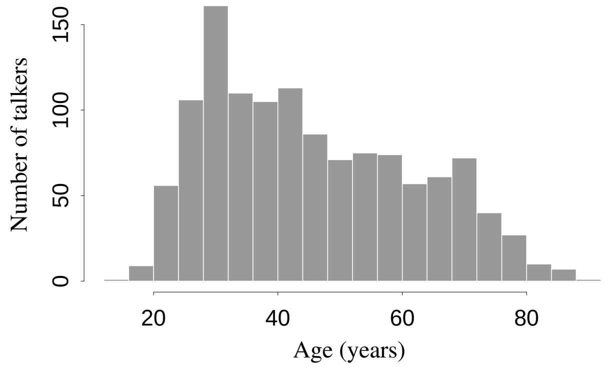

Figure 3: Histogram of talker ages. Average is 46.

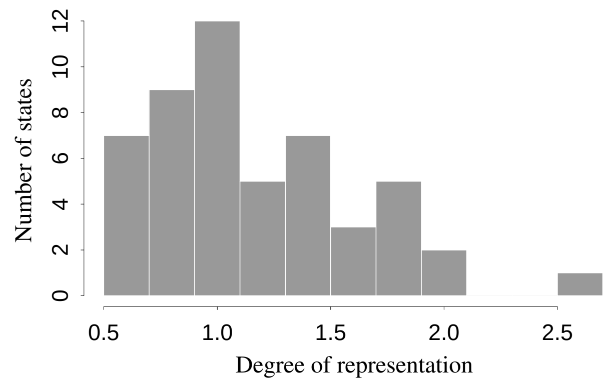

Figure 4: Comparison of geographical distribution of talker pool to distribution of population among the 50 states and Washington, DC (treated as a state for purposes of this plot). Plot is a histogram of {(state’s talkers / state’s population) / (total talkers / U. S. pop.)}. A value greater than 1 indicates that our talker pool over-represents a state compared to the overall U. S. population; less than 1 indicates an under-represented state. States with the most-deviating representation (> 2.0) are all among the five smallest states.

that the recording contained the entire word, and that the talker was reading each word “correctly”, meaning in accord with our anticipated phonemic pronunciation. Note that we *seek* a rich set of the legal *phonetic* variants on how people pronounce an underlying phoneme; thus, transcribers accepted a flapped or a fully-articulated /nt/ in a word like “winter”, but we *reject* pronunciations in which the talker set out to produce a *phoneme* sequence other than the one intended, for example, pronouncing the “t” in “mortgage” when our dictionary had transcribed it without a /t/. Transcribers rejected talkers with foreign accents, and those with high rates of truncated or mispronounced utterances. As a result, 48% of completed calls were deemed acceptable for our purposes, a much lower figure than for many other data collections because the words remained somewhat difficult, despite the dictionary filtering. The final talker pool consisted of 1341 callers, 749 female and 592 male. Average age of talkers is 46 years; Figure 3 shows a histogram. A plot comparing geographical distribution of the talker pool to that of the population is shown in Figure 4.

Given the complexity of the verification procedure, it was deemed important to verify that the transcribers shared a consistent understanding of the rejection criteria. A portion of the raw data

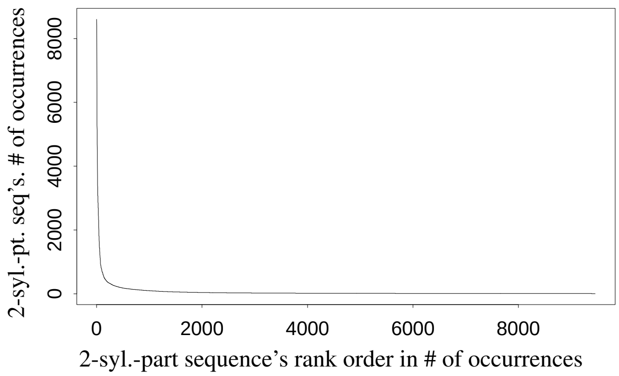

Figure 5: Distribution of frequency of occurrence of two-syllable-part sequences. The first point in the graph, (1,8603), indicates that the most-common sequence (coda /n/ followed by word boundary) occurred 8603 times. The steep drop followed by long level portion indicates that the few most-commonly-occurring contexts appear in many words (which are needed due to rarer contexts offering little or no choice of word to capture them), while the vast majority of contexts appear with a balanced distribution.

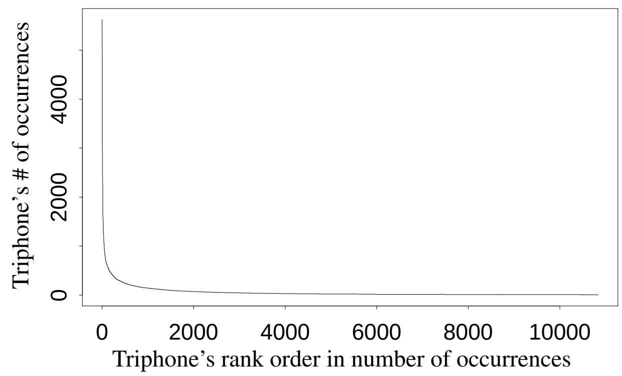

Figure 6: Distribution of frequency of occurrence of triphones. A similar explanation to that of Figure 5 applies here as well.

consisting of 1106 utterances was chosen as a transcriber-consistency-calibration set; all three transcribers agreed on acceptance or rejection on 1009 (91.2%) of these utterances. This rate of agreement per *word* compares favorably even with typical reports of agreement per *phoneme*, for example, Eisen's [7] report indicating approximately 85% overall consistency on human verification of dictionary-predicted pronunciation.

Transcribers were also asked to mark the beginning and end of speech in each utterance so that silence on either end of the word could be trimmed down to 0.3 second.

5. DATABASE STATISTICS

After verification and rejection as described above, PhoneBook consists of over 92,000 utterances totalling 23 hours of speech (based on the hand-labeled endpoints). Figures 5 and 6 show the balance in frequency of occurrence of two-syllable-part sequences and triphones, respectively. Because our goal was phonetic richness, rather than phonetic balance, we have obtained a minimum of approximately ten occurrences of very many sequences, and in so doing have obtained high representation of the relatively few most-common sequences, as can be seen in these figures.

6. SUMMARY

PhoneBook is a large phonetically-rich isolated-word telephone-speech database. In its design, we introduce a novel technique for enumerating phonemic contexts which captures several coarticulatory effects that triphones do not. We note that many types of words in a large dictionary are not suitable for inclusion in a data collection seeking specific pronunciations from naive talkers. During verification, we observed that such a data collection must contend with far greater talker and utterance rejection rates than one merely seeking any acceptable pronunciations of common words.

PhoneBook will be distributed by the Linguistic Data Consortium. While our primary interest is in developing acoustic-phonetic models and benchmarking phonetic recognition, we anticipate widespread interest in PhoneBook among speech researchers in some of the same ways that NTIMIT has been found useful.

REFERENCES

[1] Spitz, J., and the NYNEX Artificial Intelligence Speech Technology Group, "Collection and Analysis of Data from Real Users: Implications for Speech Recognition/Understanding Systems", in *Proceedings of the 1991 DARPA Speech and Natural Language Workshop*, Pacific Grove, California, February, 1991, pp. 164-169.

[2] Fisher, W., G. Doddington, and K. Goudie-Marshall, "The DARPA Speech Recognition Research Database: Specifications and Status", in *Proc. DARPA Workshop on Speech Recognition*, February, 1986, pp. 93-99.

[3] Jankowski, C., A. Kalyanswamy, S. Basson, and J. Spitz, "NTIMIT: A Phonetically Balanced, Continuous Speech, Telephone Bandwidth Speech Database", P*roceedings ICASSP 90: 1990 International Conference on Acoustics, Speech, and Signal Processing*, April, 1990, pp. 109-112.

[4] Hetherington, I., H. Leung, and V. Zue, "Toward Vocabulary-Independent Recognition of Telephone Speech", *Eurospeech 91: Proceedings of the 2nd European Conf. on Speech Communication and Technology*, Genoa, Italy, Sept., 1991, pp. 475-478.

[5] Matsuoka, T., and K. Shikano, "Robust HMM Phoneme Modeling for Different Speaking Styles", *ICASSP 91: Proceedings of the 1991 International Conf. on Acoustics, Speech, and Signal Processing*, Toronto, Ontario, Canada, May, 1991, pp. 265-268.

[6] Kassel, R., "Automating the Design of Compact Linguistic Corpora", *Proceedings of the 1994 International Conference on Spoken Language Processing (ICSLP)*, September, 1994, pp. 1827-1830.

[7] Eisen, B., "Reliability of speech segmentation and labelling at different levels of transcription," *Eurospeech '93: Proceedings of the 3rd European Conference on Speech Communication and Technology*, Berlin, Germany, September, 1993, v. 1, pp. 673-676.

ACKNOWLEDGMENTS

This work was funded largely by the Linguistic Data Consortium. We gratefully acknowledge the persevering hard work of Kelley Kertelits, Phyllis Kurtenbach and Colin O'Neal in verifying and endpointing PhoneBook; Shanmugalingam Easwaran, David Ouyang and Liliane Zreik for their extensive roles in the creation of the speech data collection platform; and David Goodine for his work on tools for endpointing and verification. We are also grateful for substantial contributions ranging from participating in formulation of this project to troubleshooting and refining algorithms and operational procedures, from Wally Anderson, Sara Basson, Henry Chang, Denise Danielson, Nancy Finnegan, Jack Godfrey, Abe Ittycheriah, Rob Kassel, Vasa Miladinov, Day Russell, Kelsey Taussig, Liang Wang and Victor Zue.