NARROW PHONETIC TRANSCRIPTION FOR DEVELOPMENT OF A LARGE VOCABULARY ISOLATED WORD RECOGNIZER

Lex Olorenshaw, Mariscela Amador, Ruxin Chen, Xavier Menendez-Pidal
Sony U.S. Research Laboratories

ABSTRACT

To build a very large vocabulary (50K) isolated word speech recognizer, speech data from over 200 native speakers of American English was recorded and manually transcribed. This paper explains the transcription method used, the motivation, and preliminary results implemented in the recognizer.

The symbol set was expanded to allow for narrow phonetic transcriptions, similar to the level of detail allowed by the International Phonetic Alphabet (IPA). By expanding the symbol set, the following was expected:

- more accurate symbolic representation of the acoustic material;
- greater consistency between transcribers;
- greater flexibility in the use of the phonetic transcriptions during training and testing of the speech recognizer;
- greater insight into the acoustic-phonetic variables that affect the performance of the speech recognizer;
- over all, to build a better recognizer.

1. INTRODUCTION

A large segment of the community involved in speech database work in the U.S. relies on the TIMIT phonetic and phonemic symbol codes for phonetic transcriptions [4] (as shown in Table 3.) Similarly, a preliminary set of data was manually transcribed at Sony using this finite set of symbols. As others have long ago noted [4,9], it was almost immediately determined that the TIMIT symbols alone would not be sufficient to account for the many detailed acoustic-phonetic variations occurring in the speech data. Our main desire was to use consistent methods in transcribing in order to provide an accurate knowledge base for training and testing the speech recognizer.

Others have attempted to achieve consistency using automated methods for transcribing instead. These methods have shown promise, but still run into difficulties due to the lack of large corpora of reliably manually transcribed and/or segmented speech data [7,11]. Our approach is to review the transcription methodology to see if improvements in reliability and consistency can be made during the manual annotation of the acoustic-phonetic database. Our first step, therefore, was to expand the symbol set to allow more detailed acoustic-phonetic transcription.

This decision was based on the fact that the acoustic-phonetic database would be the primary knowledge base upon which our large vocabulary isolated word speech recognizer technology would rely for building off-line and run-time knowledge bases. There would be no higher-level “grammar” (e.g. bigrams, trigrams) at run-time to aide the system in deciding which word had been spoken. Put simply, the oputput of the trained Hidden Markov Models (HMMs) would in turn be used directly in a search with the phonetic dictionary to find the optimal match. It has been mentioned that research in both areas, HMMs and phonetic dictionaries, will benefit from the availability of reliable, consistent acoustic-phonetic corpora [11,12,13]. Finally, there was no isolated word database suitable for our needs. (Some detailed transcription work was occurring simultaneously at ICSI for a small portion of a spontanous speech database [5], but none for isolated words.)

A speech database of approximately 130,000 isolated word tokens was recorded by over 300 speakers of American English. This resulted in approximately 25 hours of speech data. The recordings were performed in a sound-protected room at Sony U.S. Research Labs. Each of 200 speakers read a randomized list of about 600 isolated words that were displayed, one at a time, on a computer screen. Each of 100 speakers recorded about 100 words a piece. Speech was recorded using a Sennheiser Microphone HMD-410. The words were selected from the 5000 most frequent word lists from each of the following sources: the Brown Corpus[8], the British National Corpus[6] and the Switchboard Corpus[1]. Moreover, a set of less common words from the above three sources and a set of randomly chosen words from a 50,000 word dictionary were added to our final list for recording.

2. DESIGN OF EXPANDED SYMBOL SET

The expanded symbols were based on the following requirements:

1. Implement a system that is consistent for transcribers to use.
2. The notation should indicate the precise location of the acoustic change or event, that is, from left-to-right like the transitional nature of speech as displayed in waveforms and spectrograms.
3. The base phone unit and additional markings should remain clearly identifiable to the human reader, and should be easily parsable by automated methods.
4. Use characters that are standard on keyboards in the U.S., and generally avoid case distinct notation based on case.
5. To keep it simple, no attempt was made to handle duration or other intonation annotation.

2.1. Paradigm for Detailed Notation

To indicate the precise location of an acoustic change, the following paradigm was developed. First, four directional characters, greater than ‘>’, less than ‘<’, equals ‘=’ and the carat ‘^’ were introduced to separate the base phone unit from any extra notation attached to it. The mnemonic nature of these characters provides the appropriate visual cues for a quick interpretation of both the phonetic notation and the recognition of the word from the phonetic spelling.

The greater than '>' symbol may be attached to the left of the base phone unit to signify that some additional acoustic-phonetic event is occurring at the onset or during the beginning of the production of the base phone. The less than '<' symbol is the opposite. It specifies that some additional acoustic-phonetic event is occurring at the end (temporally) of the base phone. The equals '=' symbol, when used, is placed to the immediate right of the base phone. It indicates that the production of the base phone has been modified in its entirety. The carat '^' symbol indicates that an acoustic event is infixed. In other words, the event it represents is completely contained within the left-to-right bounds of the base phone. Unlike the equals sign, the carat generally indicates a brief (less than 50 msec) coarticulated event. Finally, it must also appear only to the right of a base phone unit.

This paradigm can be written in a shorthand notation, showing the flexibility that exists for creating complex phone symbols. For example, if P is a base phone unit, and X is an extra annotation symbol representing an acoustic-phonetic characteristic, then the directional characters can be used to produce the following combinations of complex phone symbols: X>P, P<X, P=X, P^X and X>P<X. Note that P, the base phone unit, is always in the root position, with X appearing in the prefix or suffix position. Furthermore, if Y and Z are different annotation symbols, then these combinations may exist: X>P<Y, X>P<Y<Z and X>P<Y=Z. The use of the directional characters also provides a syntactic division in the notation. In this way, different semantics can be assigned to the same character depending on its location with respect to a directional character. For example, 'b' as a base phone symbol is the voiced bilabial release for the stop phoneme */b/*. When in the affix position, for example with a vowel, 'b' represents breathiness, e.g. `aa<b`. In a similar way, we account for most of the detailed phonetic symbols provided by the International Phonetic Alphabet [14], and other acoustic-phonetic events that are not in the IPA. Table 1 displays a list of the acoustic-phonetic issues accounted for by this notation.

Another example is the transcription for the word "judge". A phone spelling of this word using only base phone units would be:

```
h# dcl jh ah1 dcl jh h#
```

However, it is common for the closure and release portions of the voiced alveolo-palatal affricate */jh/* to have some, if not complete, devoicing. This could be shown as

```
h# dcl<h jh=h ah1 dcl jh<h h#
```

where the 'h' character represents devoicing when in the position of an extra symbol. The first */d/* closure is devoiced in the latter portion, though it is voiced at the beginning. The first affricate release, */jh/*, is completely devoiced. The second */d/* closure is completely voiced, so no extra annotation is necessary. Finally, the last */jh/* is devoiced at the end.

Nasalization [n]	Voicing [v]	Murmur [m]
Glottalization [q]	Devoicing [h]	Air Puff [p]
Breathiness [b]	Voiced Frication [z]	Burst Quality [t]
Labialization [w]	Vless Frication [hh]	Approximation [c]
Palatalization [y]	Epenthetic Vowel [a]	No Burst/Release [u]

Table 1: Acoustic-phonetic types handled in new paradigm.

2.2. Additional Base Phone Symbols

During the transcription effort, if necessary, the introduction of a new base phone symbol was allowed. In total over 150 base phone symbols were added. This was found to be useful, not only because our phone set was never limited, but also because the transcriber was made aware of the need to account for the many acoustic changes encountered when transcribing. It allowed a transcriber to describe the actual material presented in the waveforms, spectrograms and audio signals, rather than being forced to select a poorly-matching symbol from a finite set [9]. Our preliminary transcribing exercise with only the TIMIT set proved that many times an arbitrary decision had to be made in order to select a symbol to transcribe an acoustic-phonetic event. Therefore, when human pronunciations, substantiated by evidence in waveforms and spectrograms, required the addition of a new base symbol, it could be done after being agreed upon by two expert phoneticians. More details on this topic will be addressed below.

3. SYMBOL STATISTICS

For statistical purposes, two subsets of transcribed data were selected: 1) Approximately 600 words each from 137 speakers who comprise the data set for training the Hidden Markov Models (HMMs) for the large vocabulary speech recognizer; and 2) approximately 100 words each from 104 test speakers. The latter set is made up of 2 tokens of each of the top 5154 most frequent words of English. Table 2 shows the tabulations of the transcription data for these two sets, which we refer to as Train and Test, respectively.

STATISTIC	TRAIN	TEST
Total word tokens	73,884	10,380
Total symbols (speech & silence)	916,445	118,352
Symbols per word	12.4	11.4
Total unique complex symbols	4624	1776
Total speech symbols (base phones)	513,029	68,539
Speech symbols per word	6.9	6.6
Unique base phones	169	158

Table 2: Tabulations of transcription data for two sets.

The "per word" statistics for the two sets are quite similar. Nonetheless, even the minor variance can be explained in terms of the word composition of the sets. The train set contains words that are outside of the top 5000 most frequent words. The more frequent words of English tend to be shorter, and less frequent words longer. So as more less-frequent words are introduced to the set, we expect the average number of phones per word to increase, as shown by our data.

The count of unique complex symbols is interesting since it demonstrates the ability of the paradigm for extra notation to handle an extremely large variety of acoustic-phonetic material.

An important observation for both the train and test sets is that they use significantly more base phone symbols than the widely accepted TIMIT phone symbol set. With optional primary and secondary stress markers included on the vowels and syllabic consonants, TIMIT contains 93 distinct symbols. Our set, on the other hand, allows for additional symbols for the following reasons:

1. In order to more accurately identify phone distinctions which are due to idiolect or dialectal variations.
2. To more accurately identify phones that can be interpreted as separate acoustic-phonetic events due to "stages" of the phoneme which are perceptible in the speech waveforms and spectrograms. This is the case for so-called "fricatives" which in some instances may behave acoustically as having a closure and a release.
3. To more completely identify phones whose "phoneme" covers great variability in pronunciation. This is, for example, the case for the glottal stop, which has a tremendous amount of variability at the acoustic-phonetic level.

CATEGORY	BASE PHONES
Stops & Affricates	b d g p t k dx q jh ch
Fricatives	s sh z zh f th v dh
Nasals	m n ng em en eng nx
Semivowels & Glides	l r w y hh hv el
Diphthongs & Syllabics	ey aw ay oy ow er axr
Vowels	iy ih eh ae aa ah ao uh uw ux ax ix ax-h
Silence, Closure, Pause	h# epi pau bcl dcl gcl pcl tck kcl
Stress (on vowels, etc.)	1 2

Table 3: TIMIT base phone set.

CATEGORY	NEW BASE PHONES
Variations of glottal stop	qq qh qcl qclq qqcl hqq hqh
Multiple burst releases	pp tt kk bb dd gg
Closures of fricative consonants	fcl thcl scl shcl vcl dhcl zcl zhcl
Vowel velarization/lateralization	al ol ul
R-coloring	or ar ixr
Glide loss	ee oo
R-deletion	ax_ e_ ix_ ri ra

Table 4: Categorization of additional Sony base phones.

Table 4 presents a listing of 34 more base phones that were created during our transcription project. It also provides a rough classification of the additional symbols.

4. TRANSCRIBER AGREEMENT

To examine the utility of the expanded symbol method in facilitating the creation of more uniform and consistent transcriptions across transcribers, we took a random sample of 40 speech tokens from the database. This set was transcribed by each of six transcribers and two expert phoneticians. A reference, or "correct", transcription was determined by merging information from the work of the two expert phoneticians. We anticipated that the quantity of information provided by transcribers using the detailed transcription method would allow us more flexibility in the use of the data. We explored this hypothesis further by transforming the transcribed data in various ways and re-evaluating it at each step.

4.1. Evaluation at Various Levels of Detail

The data was evaluated at four different levels to examine the performance of the transcription methodology. Conceptually, it was expected that the most detailed level of manual transcriptions would be difficult to gain high accuracy and agreement due to the many options available to the transcriber for specifying any observed acoustic-phonetic detail. However, we planned to transform the data in various ways to see if other views of the data would prove more consistent. The transformations were defined as follows:

Transformation 1: Keep all base phones, but strip away all affixation of narrow acoustic-phonetic detail. Then merge any sequentially duplicate symbols to a single symbol. The resulting base phone set size is approximately 160.

Transformation 2: Use a re-write rule method to simultaneously convert the affixation and new base phones to a slightly modified TIMIT phone set. The modifications include removing stress indicators, and keeping both stop closures and releases. The resulting base phone set size for this case is 61.

Transformation 3: Take the output from transformation #2 and collapse the phone set even further to 49 symbols. The guiding principle here is to merge voiced closures to a single phone, voiceless closures to a single phone, and merge other allophones, similar the work done with Sphinx system [2,10].

4.1.1. Many Base Phones, with Complete Detail Notation. Initially, the accuracy of each transcriber's data was calculated by comparison to the reference data. Then the agreement between transcribers was calculated. The overall accuracy for all six transcribers was 62.4%. The overall agreement between transcribers was slightly higher at 64.7%.

4.1.2. Many Base Phones, but No Extra Notation. The next evaluation was done after Transformation #1 was completed on the comparison data. In this case, the overall accuracy of transcribers jumped to 90.0%, while the agreement jumped not quite so high to 86.5%, but still significantly.

4.1.3. Reduced TIMIT Base Phones. After the second transformation on the data was complete, the calculations were run again. This time, the transcription accuracy rose to 93.3%, while the agreement continued to climb to 90.4%

4.1.4. Few Base Phones. Finally, the calculations were performed on the smaller phone set after the final transformation had been performed. Here, the accuracy dropped slightly to 92.8%, and the agreement rose slightly to 91.2%.

4.2. Evaluation Summary and Interpretation

From these results, we can clearly see the value of the enhanced transcription system as it pertains to general quality of the phonetic database. Though the phone symbol accuracy on the untransformed data seems quite low, it is clear after transformations have been completed that the additional exercise and scrutiny required by the methodology allows the transcribers to achieve greater accuracy and agreement on the base phone units. The results are summarized in Table 5.

DATA SET (Phone Set Pool)	ACCURACY	AGREEMENT
Complete Detail (4624)	62.4%	64.7%
Many Phones Only (1776)	90.0%	86.5%
Near-TIMIT Phones (61)	93.3%	90.4%
Fewest Phones (49)	92.8%	91.2%

Table 5: Accuracy and agreement values for four data sets.

5. RECOGNIZER RESULTS

The advantage of flexibility in designing the phone set for the development of an isolated word speech recognizer becomes apparent when trying to optimize system performance. Using the data from narrow phonetic transcriptions described in this paper, a speech recognizer for very large vocabulary isolated word recognition was trained. Experiments were run on various phone sets (sizes 40, 41, 42, 43 and 49) to see which would be best. Experimental models were built with triphone context dependent continuous HMMs with 3 left-to-right states. Each state has 16 Gaussian mixtures. The recognition accuracies for the five different phone sets are shown in Table 6. Two different dictionaries were used, and transformed to match the various phone sets. Further details of these results can be found in [2] and [3].

DICTIONARY:	SONY5K	CMU110K
PhoneSet1 (40 phones)	76.7%	67.0%
PhoneSet2 (41 phones)	88.2%	87.0%
PhoneSet3 (42 phones)	92.8%	86.0%
PhoneSet4 (43 phones)	92.2%	90.9%
PhoneSet5 (49 phones)	92.3%	90.2%

Table 6: Word accuracy results for experiments with five different phone sets and two dictionaries.

6. CONCLUSION

Attempts to develop automated methods for transcription are important in order to make advances in dealing with the large amount of speech data that is necessary in the development of speech recognition and related technologies. However, the results presented in this paper demonstrate the value of investing the time and resources for continued research in manual transcription methods. It is crucial that the time invested in the development of manually transcribed databases be optimized to maximize the information content of the resulting database. Limiting the transcription methods to a predetermined phone set reduces the flexibility of the database to be used in acoustic-phonetic studies related to speech recognition. On the other hand, it has been shown that a serious effort to focus on transcription methodology does allow for a high level of transcription accuracy and a high degree of agreement between transcribers. Phoneticians may never agree on the exact acoustic-phonetic content of spoken language, but a high degree of consistency can be achieved, leading to an acoustic-phonetic database with greater reliability for further research in this area.

ACKNOWLEDGMENTS

This work relies on detailed transcription work performed by the following individuals, whom we gratefully acknowledge: Fawn Bergman, Candace Cardinal, Mark Dowdy, Christine Evans, Amanda Jones, Debbie Leale, Colleen Richey, Mary Tillinghast and Jaime Wright.

REFERENCES

[1] Byrne, B., et al. 1996. *WS96 Switchboard Data Resources.* Retrieved from the World Wide Web, ftp://homer.clsp.jhu.edu/pub/swbdWS96

[2] Chen, R., Tanaka, M., Olorenshaw, L., Wu, D., Menendez-Pidal, X. 1998. Improvements on very large vocabulary isolated word recognition. *Proceedings of the 8th Sony Research Forum,* pp. 3-5, Tokyo.

[3] Chen, R., Tanaka, M., Wu, D., Olorenshaw, L., Amador, M. 1998. A four layer sharing HMM system for very large vocabulary isolated word recognition, *Proceedings of the 5th International Conference on Spoken Language Processing,* v.2, pp. 309-312.

[4] Garofolo, J., Lamel, L., Fisher, W., Fiscus, J., Pallett, D. and Dahlgren, N. 1993. *Documentation for DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus CD-ROM, NIST Speech Disc 1-1.1.* NISTIR 4930, U.S. Department of Commerce, NIST, Gaithersburg, MD.

[5] Greenberg, S., Hollenback, J., Ellis, D. 1996. Insights into spoken language gleaned from phonetic transcription of the Switchboard Corpus. *Proceedings of the 4th International Conference on Spoken Language Processing,* Philadelphia.

[6] Kilgarriff, A., et al. 1996. *British National Corpus word frequency list.* Retrieved from the World Wide Web, ftp://ftp.itri.bton.ac.uk/pub/bnc

[7] Kipp, A., Wesenick, M.-B.,Schiel, F. 1996. Automatic detection and segmentation of pronunciation variants in German speech corpora. *Proceedings of the International Conference on Spoken Language Processing*. Philadelphia.

[8] Kucera, H., et al. 1967. *Computational analysis of present-day American English.* Providence: Brown University Press.

[9] Lamel, L., Kassel, R. and Seneff, S. 1986. Speech database development: Design and analysis of the acoustic-phonetic corpus. *Proceedings of the DARPA Speech Recognition Workshop,* Report No. SAIC-86/1546, Palo Alto.

[10] Lee, K.-F., Hon, H.-W. 1989. Speaker-independent phone recognition using hidden markov models. *IEEE Transactions on Acoustics, Speech, and Signal Processing,* v.37(11), pp. 1641-1648.

[11] Schiel, F., Draxler, C., Tillman, H.G. 1997. The Bavarian archive for speech signals: Resources for the speech community. *Proceedings of the 5th European Conference on Speech Communication and Technology: Eurospeech97,* v.4, pp. 1687-1690.

[12] Wesenick, M.-B. 1996. Automatic generation of German pronunciation variants. . *Proceedings of the 4th International Conference on Spoken Language Processing*. Philadelphia.

[13] Wesenick, M.-B. and Kipp, A. 1996. Estimating the quality of phonetic transcriptions and segmentations of speech signals. *Proceedings of the International Conference on Spoken Language Processing*. Philadelphia.

[14] IPA Chart is available on the World Wide Web at http://www.arts.gla.ac.uk/IPA/ipachart.html