ALTERNATE PHONE MODELS FOR CONVERSATIONAL SPEECH

Lori Lamel and Jean-Luc Gauvain

Spoken Language Processing Group (http://www.limsi.fr/tlp)
LIMSI-CNRS, B.P. 133, 91403 Orsay cedex, France
{lamel,gauvain}@limsi.fr

ABSTRACT

This paper investigates the use of alternate phone models for the transcription of conversational telephone speech. The challenges of transcribing conversational speech are many, and recent research has focused mainly at the acoustic and linguistic levels. The focus of this work is to explore alternative ways of modeling different manners of speaking so as to better cover the observed articulatory styles and pronunciation variants. Four alternate phone sets are compared ranging from 38 to 129 units. Two of the phone sets make use of syllable-position dependent phone models. The acoustic models were trained on 2300 hours of conversational telephone speech data from the Switchboard and Fisher corpora, and experimental results are reported on the EARS Dev04 test set which contains 3 hours of speech from 36 Fisher conversations. While no one particular phone set was found to outperform the others for a majority of speakers, the best overall performance was obtained with the original 48 phone set and a reduced 38 phone set, however combining the hypotheses of the individual models reduces the word error from 17.5% (original phone set) to 16.8%.

1. INTRODUCTION

Perhaps the most notable characteristic of conversational telephone speech is its variability. Conversational speaking styles are known to vary as a function of many factors, such as the cultural and social-economic status of the parties, how well (or whether or not) they know each other, their physical, emotional and mental states, etc. Conversational speech is generally considered to be casual speech (as opposed to prepared or planned), and a lot of variability in speaking rate, loudness, and clarity of articulation can be observed even within a single conversation.

Transcribing conversational telephone speech poses many challenges at the acoustic, phonologic and linguistic levels [2]. Recent research has primarily addressed the challenges at the acoustic level (speaker normalization, channel variability, efficient speaker adaptation with small amounts of adaptation data) [3, 5, 12, 15, 16], and at the linguistic level where the primary challenge is to cope with the limited amount of language model training data [14, 17], with less focus on phonological and pronunciation modeling.

When the speech recognition community first addressed the conversational speech transcription problem in the mid 1990s (with the availability of the CallHome and SwitchBoard corpora), pronunciation modeling for this data was an active research area. This research was largely stimulated by working groups at the 1996 and 1997 Johns Hopkins Summer workshops [1]. In 1996, a working group investigated the automatic learning of word pronunciations from data, studying the most frequent word errors. The group also studied the phone deletion rates in conversational speech and found that on average one-third of the words had a phone deletion. Another working group investigated the use of a hidden speaking mode to represent systematic variations that are dependent on the syntactic structure of the word sequence. The goal was to allow different pronunciations based on the detected speaking style (sloppy vs clear vs exaggerated). During the 1997 summer workshop, a group further studied pronunciation variants in the Switchboard corpus, exploiting the ICSI hand-labeled phonetic transcriptions for 3.5 hours of data to estimate a statistical mapping between the canonical pronunciations and the realized surface forms. One of the conclusions of the 1997 group was that significant improvements in word accuracy could be obtained by modeling systematic pronunciation variations, and that estimation of the pronunciation probabilities is best done on a large corpus, with the same models as are used for recognition.

The recent availability of over 2000 hours of conversational speech in the EARS program has provided new opportunities for pronunciation modeling. Given the large differences in individual speaking styles and dialectical variations, one set of phone models may not be appropriate for all speakers and conversations. Several directions can be investigated to better take into account these differences, such as the use of alternative pronunciations for words, allowing for phone deletions, insertions, and transformations, and the use of additional lexical information such as syllable position or stress. One potential problem with using alternate pronunciations is that a large number of variants may cause confusions due to the creation of homophones. The use of pronunciation probabilities can reduce the risk of introducing confusion, and these probabilities can be estimated more reliably with the newly available data than could be done in the past. In addition, the large amount of data can result in lots of data for the most frequent phone contexts, and it may

be interesting to explore new ways for exploiting this data.

In this paper the use of four phone sets is explored in an attempt to better model different speaking styles. In addition to the standard 48 phone set used in the LIMSI CTS system, two of the alternate sets change the number of phones in a word, the first of which increases the number of phones per word, potentially better matching slow speech, and the second decreases the number of phones per word, potentially a better match for fast speech. The fourth set does not change the number of phones in the word, but introduces syllable-position dependent models for some phones which may have significantly different realizations in different syllable positions.

The remainder of this paper is as follows. The next section describes the four phone sets. Section 3 gives an overview of the LIMSI CTS system and the corpora used in these experiments, and Section 4 provides experimental results.

2. ALTERNATE PHONE SETS

Four phone sets are explored in this work, ranging in size from 38 to 129 units (phones or pseudo-phones).

The LIMSI base lexicon is represented using 48 phone symbols, with three symbols used for silence, filler words, and breath noises. The set of phones is given in [8], and includes 18 vowels /iIeE@a∧couUWYORXx|/, 6 stops /bdgptk/, 8 fricatives /sSzZfvTD/, 2 affricates /CJ/, 3 nasals /mnG/, 5 liquids and glides /wyrlh/ and 3 syllabics /LMN/.[1] A pronunciation graph is associated with each word so as to allow for alternate pronunciations, including optional phones.

The original motivation for the reduced phone set, first introduced in the LIMSI 1992 Resource Management system [9], was to permit additional sharing of contexts since the amount of training data was quite limited. This reduced phone set was later found to be efficient for faster decoding (there are fewer possible context-dependent phones, which is important when using cross word phone models) for a variety of tasks (WSJ, Broadcast News, and CTS). This phone set is also potentially better for modeling slow speech, since some of the complex phones are split into a sequence of two phones. The reduced phone set differs from the original 48 phone set as follows: the affricates /C,J/ are respectively replaced by the stop-fricative sequences /tS,dZ/; the syllabic consonants /L,M,N/ are replaced by a schwa-nasal sequence /xl,xm,xn/; the diphthongs /W,Y,O/ are replace by a vowel-glide sequence /aw,ay,cy/; the front and neutral schwas are combined together, as are the retroflex vowel and the retroflex schwa.

[1]Even though we use the term phones, the lexicon is basically phonemic, not allophonic. The use of allophones is optional, and more importantly, there is often a continuum between different the allophones of a given phoneme and the decision as to which occurred at any particular instance is subjective. By using a phonemic representation, no hard decision is imposed, and it is left to the acoustic models to represent the variants observed in the training data.

abounding	*pron1*	*pron2*
original	x b W n d | G	x b W n | G
reduced	x b a w n d x G	x b a w n x G
expanded	x -b W n- -d | G	x -b W -n | G
extended	x b W nd |G	x b W n |G
frantic	*pron1*	*pron2*
original	f r @ n t I k	f r @ n I k
reduced	f r @ n t I k	f r @ n I k
expanded	f r @ n- -t I k	f r @ -n I k
extended	fr @ nt I k	fr @ n I k

Figure 1: Example pronunciations for the words *abounding* and *frantic* using the four phone sets.

The remaining two phone sets both aim to incorporate syllable position information in the models. One of the motivations for using the syllable position is to take advantage of the large amount of training data, using multiple syllable-position dependent models for data that were previously modeled by only a single context. In the expanded phone set, the number of phones in the word is unchanged, but certain frequent phones which may have different realizations in different syllable positions are differentiated. In the extended phone set, some selected phone sequences are mapped into a single unit, in an attempt to better model heavily coarticulated and fast speech.

The syllabification algorithm uses the maximum onset principle with stress resyllabification [7]. That is, if there are multiple possible locations for a syllable boundary in a consonant string, the syllable boundary is placed just before the longest permissible syllable-initial sequence. For example, in the word *ashtray* /'@S-tre/, /Str/ is not a permissible syllable onset sequence, so the syllable boundary is placed before the /t/. In the word *astray* /x-str'e/, the syllable boundary (denoted by -) is place before the /s/. The syllabification algorithm differentiates hard syllable boundaries from ambisyllabic ones using stress resyllabification. When the lexical stress is falling, the left-most consonant in the syllable-onset is marked as being ambisyllabic, as opposed to being the first consonant following the syllable boundary. For example, the syllabified pronunciation for the word *pretty* is /pr'I<ti/, where the < sign indicates a soft syllable boundary, i.e., the /t/ is ambisyllabic. This is in agreement with the common realization of the /t/ as a flap.

The expanded phone set contains a total of 101 phone-like units. The phones for the stops /bdgptk/, the fricatives /sSzfvTD/, the nasals /mn/, the liquids /lr/ and the vowels /ie@WY∧OouRxX/ were expanded to have separate symbols when occurring in (word-internal) syllable-initial and syllable-final positions. For this phone set, all ambisyllabic markers were treated as hard syllable boundaries. As an example, this expanded phone set uses different symbols to represent the /t/'s in *outrage*, *nitrate* and *night*.

The extended phone set is also syllable-position dependent but aims at capturing some of the strong coarticulations and reductions found in casual speech. Fast speakers tend to poorly articulate unstressed syllables (and sometimes even skip them completely), particularly in long words with sequences of unstressed syllables. This extended representation maps some frequent phone sequences into pseudo-phones which can represent a consonant cluster, a vowel-liquid sequence or a vowel-nasal sequence. The motivation for the pseudo-phones is that sometimes when the coarticulation is strong it is difficult to temporally segment the two phones. Some well-known examples are the retroflex-color of the /a/ in the word *arm* and the /d/ in *dress* and the nasalization of the vowel in the word *can't*. The consonant sequences represented by a single pseudo-phone are: /sts, nts, ns, ndz, nz, nd, n<d, nt, n<t, dz, mps, ms, mp, m<p, Ggz, Gz, dr, tr, gr, pl, kl, kr, fr, fl, Tr, Dr, st, ts, G<g, Gg, Gks, G<k, Gk/, where < indicates here that the following consonant is ambisyllabic. Other pseudo-phones represent the vowel-/r/ and vowel-/l/ sequences for the vowels /i,I,e,E,@,a,c,o,u,U,^,x/, and shortvowel-nasal and schwa-nasal sequences. The sequences /sxz, yur, yUr, yu , yU/ are also represented by a pseudo-phone.

Figure 1 gives the representations using the four phone sets for the words *abounding* and *frantic*.

3. CORPUS AND SYSTEM OVERVIEW

The acoustic and language models used in these experiments were trained on the LDC conversational speech data including 430 hours from the SwitchBoard corpus (SWB) [6] and 1864 hours from the Fisher corpus. There is a total of about 2300 hours of speech from about 15k conversations, with 55% of the data coming from female speakers. The experimental results reported in this paper were obtained on the EARS Dev04 test set which includes 36 Fisher conversations for a total of 3 hours of speech.

The acoustic models are tied-state position-dependent crossword triphones. The state tying is obtained with divisive decision tree based clustering algorithm and a set of questions related to the phone context. Two sets of gender-dependent models are built after dividing the training data into the gender specific subsets. More details about the way these models are estimated as well as a description of the acoustic front-end (PLP cepstrum with VTLN) can be found in [5, 13]. To limit the computational resources used for these experiments the models were ML trained, i.e. we did not use any form of discriminative estimation.

The decoder uses multiple decoding passes. Each pass generates word lattices with a trigram language model. The lattices are then expanded with a fourgram language model, prior to performing a consensus decoding [11] with pronunciation probabilities. Between each decoding pass unsupervised acoustic model adaptation using the MLLR technique [10] is carried out.

The trigram and fourgram language models were obtained by interpolating models trained on various data sets, of which the most important are the transcriptions of the CTS training data (27M words), and transcriptions of broadcast news (370M words). The fourgram backoff LM is also interpolated with a neural network LM trained on only the transcriptions of the CTS data [14].

The recognizer vocabulary includes 49959 words selected from the same data sets so as to maximize the coverage on a development set from the Fisher corpus. The pronunciation dictionary has a total of 59381 phone transcriptions for the 49959 words. The basic pronunciations are taken from the LIMSI American English lexicon, for which the most frequent inflected forms have been verified to provide more systematic pronunciations. The pronunciation probabilities are estimated from the observed frequencies in the training data resulting from forced alignment, with a smoothing for unobserved pronunciations.

For each phone set studied in this work, the following resources had to be developed: a phone-set specific pronunciation dictionary with pronunciation probabilities; a set of questions for the state tying of the acoustic models; and the acoustic models, which requires several iterations of segmenting the entire 2300 hour corpus and estimating the parameters of the new models.

The four set of gender dependent acoustic models built for these experiments are summarized on Table 1. All the model sets have about 30k tied states, except for the models based on the extended phone set which have around 50k states. For all models there are about 32 Gaussians per state.

Set	*#phones*	*#contexts*	*#tied states*
Original	48	38k	30k
Reduced	38	28k	30k
Expanded	101	52k	30k
Extended	129	58k	50k

Table 1: Characteristics of the phone-set specific acoustic models for the standard 48 phone set, the reduced 38 phone set, the expanded 101 phone set and the extended 129 phone set.

4. EXPERIMENTAL RESULTS

The recognition results on the EARS Dev04 data using the models described in the previous section are given in Table 2. The first three decodes use only the original phone set and result in a 17.5% word error rate after two passes of MLLR adaptation using 2 regression classes (speech and non speech) and 4 phonemic regression classes, respectively.

The word error rates obtained with the three other model sets are given in the second part of the table. These decodes are also preceded by a 4 class MLLR adaptation with the same second pass hypothesis as was used for the original phone set decode. Comparable results are obtained for the three model set, even though the extended set appears to have

Decode	*WER*
Unadapted fast decode	23.4
2 class adaptation	17.8
4 class adaptation	17.5
Reduced set, 4cl adapt.	17.3
Expanded set, 4cl adapt.	17.6
Extended set, 4cl adapt.	17.8
Combination	16.8

Table 2: Decoding results using the 4 phone sets.

the highest error rate (17.8%).

Looking at the results speaker by speaker we found that no one single phone set performs best for the majority of speakers. The standard, reduced and expanded phone sets each give the best performance for 20.5 of the speakers in the development data (ties were counted as 0.5), whereas the extended phone set gives the best results for 10.5 speakers. We listened to portions of the data from the speakers who had the lowest word error rates with the extended phone set, and our impression is that most of these speakers have a casual speaking style, with a tendency of slurring some of their words. We also listened to some of the data for speakers who had at least a 1% absolute improvement with an alternate phone set over the standard one. Our expectations were that the reduced phone set would favor slow speakers, but we did not find any indications that speakers with slower average speaking rates have better recognition accuracies with this phone set.

If the models outputs are combined with Rover [4] the word error rate is reduced to 16.8% showing that these model sets are somewhat complementary. Comparing the combined result to the original results at the speaker level, we found that the gain can be quite large for some speakers (4% absolute, up to 30% relative) but there is no large loss for any speaker. We also observed that this setup does not seem to help bad speakers more than good ones, and conversely do not improve the most on good speakers. Using an global estimate of the speaking rate in words per minute, there was no notable improvement for speakers with slow or fast speech.

In order to make the model combination more appealing, the same experiments carried out by replacing the regular decodes for the alternate phone sets by lattice decodes, using the lattices generated in the third decoding pass. Even though the lattice decode takes only 1/10th of the CPU time needed for the regular decode, the combined result is almost identical (16.9%).

5. CONCLUSIONS

This has described recent experiments assessing four alternate phone sets for the transcription of conversational telephone speech. The alternate phone sets ranged in size from 38 to 129 units. While no one particular phone set was found to outperform the others for a majority of speakers, the best overall performance was obtained with the original 48 phone set and the reduced 38 phone set. This advantage may be real, or may simply be due to the more extensive experience we have with these phone sets, since we have built many more acoustic models sets with them.

REFERENCES

[1] The Center for Language and Speech Processing at Johns Hopkins University Summer Workshops, 1996, 1997. http://www.clsp.jhu.edu/ws96/. http://www.clsp.jhu.edu/ws97/.

[2] C.S. Culhane & B.J. Wheatley, "Hub-5 Conversational Speech Recognition," *Proc. DARPA Speech Recognition Workshop,* Harriman, February 1996.

[3] G. Evermann et al., "Development of the 2003 CU-HTK Conversational Telephone Speech Transcription System," *Proc. IEEE ICASSP'04*, **1**:261-264, Montreal, May 2004.

[4] J. G. Fiscus, "A post-processing system to yield reduced word error rates: Recognizer output voting error reduction (ROVER)," *Proc. ASRU'97*, 347354, Santa Barbara, December 1997.

[5] J.L. Gauvain et al., "Conversational Telephone Speech Recognition," *Proc. ICASSP'03*, April 2003.

[6] J.J. Godfrey, E.C. Holliman & J. McDaniel, "Switchboard: Telephone speech corpus for research and development," *ICASSP'92*, **1**:517-520, San Francisco, March 1992.

[7] D. Kahn, "Syllable-based generalizations in English phonology," PhD dissertation, MIT, 1976.

[8] L. Lamel & G. Adda, "On designing pronunciation lexicons for large vocabulary, continuous speech recognition," *Proc. ICSLP'96*, **I**:6-9, Philadelphia, October 1996.

[9] L. Lamel & J.L. Gauvain, "Continuous speech recognition at LIMSI," *Proc. Final review of the DARPA Artificial Neural Network Technology Speech Program*, Stanford, 59-64, Sept. 1992.

[10] C.J. Leggetter & P.C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," *Comp. Speech & Lang.*, **9**(2):171-185, 1995.

[11] L. Mangu, E. Brill & A. Stolke, "Finding Consensus Among Words: Lattice-Based Word Error Minimization," *Eurospeeech'99*, 495-498, Budapest, Sep. 1999.

[12] S. Matsoukas et al., "BBN CTS English System, in DARPA RT-03 Workshop, Boston, May 2003, http://www.nist.gov/speech/tests/rt/rt2003/spring/presentations/cts-combined-sm-ok-v14.pdf

[13] R. Schwartz et al., "Speech Recognition in Multiple Language and Domains: The 2003 BBN/LIMSI EARS System," *Proc. IEEE ICASSP'04*, **III**:753-756, May 2004.

[14] H. Schwenk & J.L. Gauvain, "Neural Network Language Models for Conversational Speech Recognition," *Proc. ICSLP'04*, to appear, October 2004.

[15] H. Soltau et al., "The 2003 ISL Rich Transcription System for Conversational Telephony Speech," *Proc. IEEE ICASSP'04*, **I**:773-776, May 2004.

[16] A. Stolcke et al., "Speech-to-text research at SRI-ICSI-UW", in DARPA RT-03 Workshop, Boston, May 2003, http://www.nist.gov/speech/tests/rt/rt2003/spring/presentations/sri+-rt03-stt.pdf

[17] W. Wang, A. Stolcke, & M. P. Harper, "The Use of a Linguistically Motivated Language Model in Conversational Speech Recognition," *Proc. IEEE ICASSP'04*, **1**:261-264, Montreal, May 2004.