FABRICATING CONVERSATIONAL SPEECH DATA WITH ACOUSTIC MODELS: A PROGRAM TO EXAMINE MODEL-DATA MISMATCH

Don McAllaster, Larry Gillick, Francesco Scattone, Mike Newman

Dragon Systems, Inc.
320 Nevada Street
Newton, MA 02160

ABSTRACT

We present a study of data simulated using acoustic models trained on Switchboard data, and then recognized using various Switchboard-trained acoustic models. When we recognize real Switchboard conversations, simple development models give a word error rate (WER) of about 47 percent. If instead we simulate the speech data using word transcriptions of the conversation, obtaining the pronunciations for the words from our recognition dictionary, the WER drops by a factor of five to ten. In a third type of experiment, we use human-generated phonetic transcripts to fabricate data that more realistically represents conversational speech, and obtain WERs in the low 40's, rates that are fairly similar to those seen in actual speech data.

Taken as a whole, these and other experiments we describe in the paper suggest that there is a substantial mismatch between real speech and the combination of our acoustic models and the pronunciations in our recognition dictionary. The use of simulation appears to be a promising tool in our efforts to understand and reduce the size of this mismatch, and may prove to be a generally valuable diagnostic in speech recognition research .

1. MOTIVATION

In recent years, steady progress has been made in automatic recognition of conversational telephone speech [1]. Nevertheless, state-of-the-art systems, running at hundreds of times real time, using hundreds of megabytes of memory, still have word error rates of more than 30%. How much improvement can be expected? For example, can we achieve a 10% error rate recognizing conversational telephone speech?

In this paper, we seek to shed some light on these matters by simulating speech data from speech models, and then exploring the performance of our standard speech recognition algorithms when applied to this data. The great merit of simulated data is that we understand, and can control, the probability mechanism that produces it. The use of simulated data in probing the strengths and weaknesses of pattern recognition algorithms is standard practice in the mainstream statistical literature and is, perhaps, not so common in speech recognition circles as it should be. A subsidiary goal of this paper, therefore, is to provide an example of the fruitful use of this sort of technique.

Our focus in the experiments we report here is on acoustic modeling and on pronunciations. Although signal processing and language modeling are undoubtedly important components in speech recognition, and progress in these areas continues to be made, we believe that acoustic and pronunciation modeling provide the most fertile fields for improvement.

A primary source of concern with our present modeling techniques is simply that real speech data may not be adequately described by our acoustic models. By generating data from the acoustic models, we can, in essence, eliminate the problem of "mismatch". What will happen when we try to recognize such data? Suppose that the error rate remains high. This would suggest that the acoustic states of conversational speech, as captured by our training procedure, are inherently poorly separated and confusable. If the error rate is near zero, that would suggest that there is a serious problem of mismatch between model and data. The experiments we have done will suggest that the mismatch problem is a sizable one, and that, in particular, the mismatch between the pronunciations in our standard lexicons and those that are actually used by people in conversation may be the key to the puzzle.

In this paper, Section 2 gives an overview of our two main schemes for simulating data, along with a description of the test set, and the acoustic and language models used. Section 3 goes on to discuss a series of experiments with simulated and real data, and Section 4 draws some conclusions.

2. SIMULATING DATA

In our experiments we use two data simulation schemes. In the first, we generate data using the pronunciations in our recognition dictionary, while the second makes use of hand-labeled phonetic transcriptions.

The experiments reported in this paper are based on the "test-ws96dev-i" devtest, used in the 1996 and 1997 summer workshops at Johns Hopkins [2], whether it be real data or simulated. While this test is rather small – 6 two-sided conversations, lasting 23 minutes, and composed of 4700 words total – ICSI (the International Computer Science Institute at the University of California at Berkeley) has made hand-labeled and time-marked word and phonetic transcripts of it [3]. We use these invaluable transcriptions to determine the phonemes actually uttered in the conversations. Of course, humans are not infallible; in particular, two experienced transcribers will disagree over 20% or so of the phoneme tokens in this difficult material [3]. Nevertheless, these being the best available phonetic transcriptions, we use them.

2.1. Simulation from Phonetic Transcript

One data simulation scheme begins with ICSI's devtest phonetic transcripts. Using a transliteration table, which maps each of the ICSI phonemes to one or two of Dragon's Switchboard phonemes, we convert ICSI's phonemes into ours, and decompose the resulting phoneme string into a sequence of context-dependent phoneme states. For each state, we randomly choose a

component from the state's mixture model, based on the mixture weights, and probabilistically generate a frame from the component. We determine how many frames to generate for each phoneme state by simulating an observation from the duration model for the state.

2.2. Simulation from Dictionary

Our other simulation method determines the triphone sequence differently, by assuming (incorrectly) that people pronounce words exactly as in our recognition dictionary. We take word transcriptions of the ICSI devtest, and look up pronunciations for the words in our dictionary, randomly choosing among multiple pronunciations for a word as necessary. We reduce the selected string of pronunciations into a sequence of context-dependent phoneme states, and proceed exactly as described above.

This data simulation scheme is likely to produce less realistic data than simulating from phonetic transcript. As we shall see below, the pronunciations used in conversational speech are far more varied than recorded in our recognition dictionary, and by requiring that words be pronounced according to our dictionary, we may significantly understate the phonological variety of speech.

2.3. Acoustic and Language Models

We train an acoustic model from a Viterbi alignment of 60 hours of Switchboard data, and also train two 30-hour models on data divided such that the two sets are gender-balanced, and share no speaker.

The language model is constructed with all the bigrams and unigrams in the three million word Callhome and Switchboard training sets, smoothed by absolute discounting; the vocabulary is constructed by taking all 28000 distinct words found. Our recognition dictionary has about 32000 pronunciations for these words; about 3500 of the words have more than one pronunciation. All alternate pronunciations are treated as equally probable by the recognizer.

3. EXPERIMENTS

We present two series of experiments: comparing recognition of real and simulated data, and exploring the failure to improve recognition of real data by augmenting the pronunciations in the recognition dictionary.

3.1. Comparing Simulated and Real Data

In this experiment, we recognize both real and fabricated speech data, using the three acoustic models described above. The fabricated data is generated using the first (AM1) of the 30-hour acoustic models. We see in Table 1 that for real data, the two 30-hour models produce very similar WERs, while the 60-hour model is about 2 percentage points better. This is a typical result; it shows that the two 30-hour sets, while yielding comparable recognition results, contain at least partly complementary information.

Recognition of the speech simulated from dictionary gives a very different picture. When we recognize with the same acoustic models that we used to generate the data, the error rate drops below 5%. This situation corresponds to recognition with models trained on an infinite amount of data; by construction, the data complies perfectly with the probability assumptions of the model. When we recognize with models trained on completely disjoint data (AM2), the error rate doubles, but still hovers near 10%. This behavior illustrates the fact that the generating and recognizing models, trained on different data, disagree. The 60-hour models have seen AM1's training data, but are led in a somewhat different direction by AM2's: the result is an error rate between that of the two 30-hour models.

Test Set	30hr AM1 WER (%)	30hr AM2 WER (%)	60hr AM WER (%)
Real Data	48.2	48.8	46.3
Data simulated from dictionary	4.3	10.8	8.4
Data simulated from phonetic transcription	41.3	43.9	41.4

Table 1: Baseline WER and WERs when recognizing data simulated with AM1, along with either a dictionary, or with phonetic transcripts.

We can take some encouragement from these results. The acoustic models appear to be sharp enough that simulated data is recognized incorrectly five to ten times less often than real data. In other words, while you might assign some of the mistakes in recognition of real speech to the confusability of our models, most of the errors appear to be due to something else!

So if we use our recognition dictionary to choose pronunciations for words in a transcript, and generate speech data from the pronunciations that complies with the probability assumptions of our acoustic model, we can get impressively good recognition results. But what happens when we relax the requirement that data be generated from pronunciations in our recognition dictionary?

In the third entry of Table 1, the data is fabricated using the 30-hour acoustic model 1, along with the ICSI phonetic transcripts, without recourse to the pronunciations in the recognition dictionary. Word error rates are only a little better than those obtained when recognizing real data. Even recognizing with the same acoustic models that generated the data (AM1) – in other words, with acoustic models that perfectly represent the triphones used – makes only a small difference.

This contrast is striking. When we force words (through the simulation process) to be pronounced according to our recognition dictionary, we get astoundingly good recognition, but when words are simulated with pronunciations that more fairly represent the diversity found in conversation, the error rate is nearly as high as for real speech. Put most provocatively, the variant and reduced pronunciation of casual speech accounts for most of the errors made by this recognition system.

That words are pronounced in unexpected ways in conversational speech is not, itself, unexpected. In fact, the manner in which we train our acoustic models works to reduce the impact of unexpected pronunciations. We train from alignments in which each frame of training data is mapped to a phoneme state; the states to which we align are determined from the word transcription by the recognition dictionary. When the word is pronounced according to the recognition dictionary, then the model for that triphone stands a chance of being trained on the right data. When it's pronounced differently (if we trained on this test set, that would

happen about half the time), then the alignment will be incorrect, assigning frames from the wrong triphone. By using decision tree clustering, and multiple components in each mixture, the models can deal to some extent with this sort of misassignment, but end up being more diffuse, and needing more components, than they might otherwise. Furthermore, since the WER on data that matches the triphone models perfectly is almost as bad as for real speech, adding components proves not to be as robust against missing pronunciations in training as might be hoped.

3.2. Dictionary Expansion – Simulated Data

If our recognition dictionary lacks useful pronunciations, then perhaps we can try to improve it by adding some. Note that others (eg, [5]) have also done this with real data; by and large they have seen only small improvements in performance. We add every pronunciation we find in the phonetic transcripts, even if it occurs only once. We note that there are about 4700 tokens in the test data, amounting to 900 distinct words. Only 47% of the tokens are pronounced as in our dictionary. About 650 words are pronounced only one way in the test data, while *the* has 36 different pronunciations, according to the transcripts. Only about a quarter of the 2100 test data pronunciations are in our dictionary, so we end up adding 1500 new ones to create the "base + test" dictionary of Table 2.

Dictionary	30hr AM1 WER (%)	30hr AM2 WER (%)	60hr AM WER (%)
base	41.3	43.9	41.4
base + test	23.9	33.5	29.8
base + train	50.6	50.3	48.2
base + test + train	30.4	40.2	35.7

Table 2: Simulated data, recognized using baseline and augmented dictionaries. Data is simulated with the 30hr AM1, using the ICSI phonetic transcriptions to determine the triphones.

Note that while all of the acoustic models experience improved recognition, AM1 improves the most; the better the acoustic model matches the data, the greater the benefit from having an augmented dictionary. This is another instance of a "perfect" dictionary, as with data simulated from dictionary in Table 1. In this case, however, instead of the error rate dropping to 5% or 10%, it goes down to 20% or 30%. The difference appears to be confusability among the prons: there are many more homonyms and near homonyms in the "base + test" dictionary than in the base dictionary alone. For example, in our base dictionary, only one pronunciation is associated with as many as five different words: *sons, son's, sons', suns,* and *sun's*; no pair of words shares more than two pronunciations. By contrast, the "base + test" dictionary has 38 pronunciations associated with 5 or more words, headed by *schwa*, which is a pronunciation for 27 different words. Nineteen word pairs share three or more pronunciations; the most confusable pair is *the* and *to*, which have 7 pronunciations in common.

Of course, it is cheating to look into the test data to discover new prons. Suppose we gather pronunciations from a different set of phonetically transcribed data: the "train-ws96-i" set, also produced by ICSI and used in the 1996 and 1997 summer workshops at Johns Hopkins. This data has about 10000 word tokens, of which 1500 are distinct, pronounced 3400 ways. About 500 of these words are shared with the test data; of these shared words, about 700 word/pron pairs are held in common, and 1400 are unique to the training data. For example, *the* has 38 pronunciations in the training data; only half of these are observed in the test set. In addition, the training data has 1000 words (with 1300 prons) that don't occur in the test data. After adding all the training pronunciations to our dictionary, about 71% of the word tokens in the test set are pronounced as in the dictionary, up from 47% before expansion.

The "base + train" entry in Table 2 gives recognition results after we have added these training pronunciations to our base dictionary It is noteworthy that all of the acoustic models yield degraded performance with this dictionary. We have evidently added too much confusability, and too few of the pronunciations that do occur in the test data. It also gives some notion of the futility of simply adding pronunciations en masse: it is all too easy to make recognition worse.

Recognition results when both the test and training pronunciations are added are listed on the "base + test + train" line of Table 2. All the acoustic models experience improved results compared to the "base" recognition, despite the confusability added by the extra pronunciations and inevitable homonyms (the phoneme *schwa* is a pronunciation for 35 different words; 79 pronunciations have 5 or more homonyms). For the simulated data, it appears that is possible to improve recognition by adding pronunciations to the dictionary – provided that "enough" of the pertinent ones are added relative to the increased confusability.

We can see the effects of confusability in these results by examining the kinds of errors we are making in, as reported in Table 3. This data is generated with AM1 along with the phonetic transcripts, and recognized using AM2. Note that many of the added pronunciations are very short, corresponding to common words, and so tend to be inserted frequently. In general, adding prons decreases the number of deletions, but increases the insertion rate. Adding the more pertinent test pronunciations decreases substitutions, while adding the training prons tends to increase them.

Dictionary	Total	Insertions	Deletions	Substitutions
base	2063	99	710	1254
base + test	1577	184	360	1033
base + train	2364	346	376	1642
base + test + train	1891	236	359	1296

Table 3: Breakdown of errors by type, for synthetic data generated by AM1, and recognized using AM2 and the baseline and augmented dictionaries.

3.3. Dictionary Expansion – Real Data

Because adding the test pronunciations to the lexicon appeared always to improve recognition performance, even when many other misleading prons are also added, we wanted to repeat these experiments with real data instead of phonetically-simulated data. The results are listed in Table 4.
We see that in all cases, adding more pronunciations to the recognition dictionary seriously degrades performance. Even when we cheat, and add only the pronunciations that we know will occur in the test set, recognition still gets worse. This is in sharp contrast to the situation with simulated data: for example, when we add the test prons to the dictionary and recognize with AM2, the

Dictionary	30hr AM1 WER (%)	30hr AM2 WER (%)	60hr AM WER (%)
base	48.2	48.8	46.3
base + test	58.6	60.8	58.5
base + train	64.3	65.7	63.1
base + test + train	65.3	66.8	65.5

Table 4: Real data, recognized using baseline and augmented dictionaries.

WER for simulated data drops from 43.9% to 33.5%, whereas it increases from 48.8% to 60.8% for real data.

Analysis of the errors made (Table 5) shows a pattern broadly similar to synthetic data, although to a degree less favorable to a low WER. Adding pronunciations tends to increase insertions and decrease deletions, just as with synthetic data, but with real data insertions increase more and deletions decrease less. Real data differs from simulated data, in that the number of substitutions increases whenever pronunciations are added.

Dictionary	Total	Insertions	Deletions	Substitutions
base	2296	323	461	1512
base + test	2861	616	334	1911
base + train	3091	729	297	2065

Table 5: Breakdown of errors by type, for real data recognized using baseline and augmented dictionaries.

We believe this discrepancy is more evidence of the mismatch between real speech and our acoustic models, or, equivalently, the difference between real and simulated speech. Consider the speech from which a model is trained. The speech contains frames from the right triphones, and many wrong ones. We build a model from the data, averaging the frames from right and wrong triphones into a few gaussians. When we build models from disjoint data, and use one model to generate data and the other to recognize, we find simulated data is easier to recognize than real data. That is because the recognition algorithm and simulated data are both built upon the compromises implicit in model building. Models built upon two sets of disjoint data can be made arbitrarily similar by training from more and more data, assuming that speech exhibits a finite amount of diversity. Despite this convergence, a 30-hour model is still surprised by phenomena in real speech, more surprised than by data generated from the other model.

Both simulated and real speech remain susceptible to confusability from exact homonyms in the dictionary, but since simulated speech is a better match to acoustic models (even ones trained on different data), it is less vulnerable to near homonyms. That is why speech fabricated from a model can take better advantage of added pronunciations.

We can see this effect at work when we compare the error rate for words pronounced according to our dictionary with words pronounced differently (Table 6). We consider the non-cheating case, where we generate data with AM1, and recognize with AM2. For each word token in the correct transcript, we record whether it is pronounced according to the recognition dictionary, and whether it was recognized correctly, thus compiling in-dictionary and out-of-dictionary error rates. Note that these statistics are smaller than the word error rate, since it does not account for errors due to insertion.

Data Source	Error rate: prons in dictionary	Error rate: prons out of dictionary	Error rate: overall
real data (base)	35.4	47.4	41.8
data simulated from phonetic transcript (base)	24.1	57.3	41.7
real data (base + train)	46.3	59.6	50.2
data simulated from phonetic transcript (base + train)	34.5	63.5	42.9

Table 6: Error rates for words in reference transcripts, broken down by whether their pronunciations are in the recognition dictionary.

As might be hoped, word tokens pronounced according to the dictionary are more likely to be recognized correctly than tokens pronounced in an unexpected way. But the difference between the error rates is smaller for real than for synthetic data, since the models do not match up so well with real speech as with simulated. Having the just the right pron is less important for real speech, and similarly, lacking the right pron is less costly.

4. CONCLUSION

We have outlined an avenue of investigation using data fabricated from acoustic models. Data simulated from dictionary pronunciations tend to WERs of 5% to 10%. When the data is simulated from phonetic transcriptions, WERs rise into the 40%s; when we add dictionary pronunciations, we see a decrease in the error rate for simulated data, so long as "enough" correct prons (the ones that occur in the test set) are included. Real data, on the other hand, always gets worse recognition results, at least when the dictionary is augmented in this unconstrained way, and the recognizer does not apply unigram or bigram probabilities to alternate pronunciations. We believe this discrepancy is due to a mismatch between real speech and the models built from it. At least part of this mismatch is due to the extremely varied pronunciations found in conversational speech, and the way we train our models. Reducing this mismatch will be vital to continued progress in automatic recognition of conversational speech.

References

1. LVCSR Hub 5 Workshop Proceedings, 1995, 1996, 1997.
2. 1996 CLSP/JHU Workshop on Innovative Techniques for Large Vocabulary Continuous Speech Recognition, July 15 - August 23, 1996, Center for Language and Speech Processing, Johns Hopkins University, Baltimore, MD 21218
3. Steven Greenberg, Joy Hollenback, Dan Ellis, "Insights into spoken language gleaned from phonetic transcription of the Switchboard corpus", Proceedings Addendum, ICSLP '96, pp. 24-27
4. B. Peskin et al. "Progress in Recognizing Conversational Telephone Speech," Proc. ICASSP-97, Munich, April 1997.
5. B. Byrne et al., "Pronunciation Modeling for Conversational Speech Recognition: A Status Report from WS97," IEEE ASRU Workshop, Santa Barbara, December 1997.