

# STUDIES WITH FABRICATED SWITCHBOARD DATA: EXPLORING SOURCES OF MODEL-DATA MISMATCH

*Don McAllaster, Larry Gillick, Francesco Scattoni, 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. The Switchboard-trained models yield word error rates of about 47 percent, on real Switchboard conversations. When data is simulated using the acoustic models, but in a way that insures that the pronunciations in our recognition dictionary are “perfect”, the WER drops by nearly a factor of five. If instead we use hand-labeled phonetic transcriptions to fabricate data that more realistically represents the way words are pronounced – rendering our recognition pronunciations imperfect – we 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 data and our speech models. The use of simulation in speech recognition research appears to be a promising tool in our efforts to understand and reduce the size of this mismatch.

## 1. MOTIVATION

Why is the error rate so high on conversational speech? How much can it be improved? For example, is a 10% error rate on conversational speech achievable? In this paper, we seek to shed some light on these matters through the vehicle of 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 lies in the fact that the underlying probability mechanism that produced it is known and, indeed, controllable. The use of simulated data in probing the strengths and weaknesses of pattern recognition algorithms is common, even 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. All experiments will be based on Dragon’s standard front end, which involves 24 IMELDA parameters derived from an original set of 44 parameters (spectral and cepstral parameters together with first and second cepstral differences), and on the use of a standard bigram language model. We note that it has been notoriously difficult to make substantial improvements in the language model in speech recognition, and, on the other hand, that standard signal processing techniques are good enough that it has been possible to achieve error rates around 10% [1] on some large vocabulary recognition tasks. It therefore appears that the greatest prospects for improvement must lie in the area of acoustic and pronunciation modeling, and we focus our attention there.

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 simulating data from the acoustic models, we can, in essence, eliminate the problem of “mismatch”. What will happen when we try to recognize such data? Will the error rate be near zero, or instead, will it turn out that the error rate will still be high? The latter consequence would suggest that the acoustic states are not well separated in acoustic space, while the former 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 to be 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 our recognition dictionary, while the second makes use of hand-labeled phonetic transcriptions as the starting point.

All the results presented 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. This test is rather small: 6 two-sided conversations, lasting 23 minutes, and composed of 4700 words, but ICSI (the International Computer Science Institute at the University of California at Berkeley) has made hand-labeled and time-marked word and phonetic transcriptions of it [3]. We use these invaluable transcriptions in the experiments described below.

### 2.1. Simulation from Dictionary

One data simulation method begins with word transcriptions of the test data. We took these transcriptions, and looked up pronunciations for the words in our recognition dictionary. If the dictionary had multiple pronunciations for the word, we chose one randomly. We decomposed the selected pronunciations for the words into a sequence of triphones, and then for each state in each triphone, randomly chose a component (based on the mixture weights) from the state’s mixture model. We then generated a sequence of frames for the triphone state, taking into account the component’s mean and variance, and determining the number of frames generated with the

mean and variance of the state’s duration model. In these experiments, the recognition dictionary is perfect, since our test data is generated via the pronunciations in the dictionary: the words are constrained to be “pronounced” (by the simulation) exactly as the dictionary says.

## 2.2. Simulation from Phonetic Transcription

Our other data simulation scheme determined the triphone sequence differently. Rather than starting with the word transcription and our recognition dictionary, we instead began with ICSI’s phonetic transcriptions of the same conversations. We stripped the diacritical marks from the transcriptions, and transliterated each of the ICSI phonemes to one or two of the phonemes used in Dragon’s Switchboard work. The resulting triphones were fed into the simulation process as above. This set of experiments results in more realistic data than the first set, as the triphones that are used for simulation are the ones that were actually used by the speakers (up to transcription and transliteration errors), and not merely the ones which happened to appear in the dictionary pronunciations for the words that were uttered. As in the first set of experiments, we also used the duration model for the triphone state to determine how many frames of data to use; we did not use the time marks in the phonetic transcriptions for this purpose.

## 2.3. Acoustic and Language Models

Our initial acoustic models are trained on 60 hours of Switchboard data. We divide the data into two 30 hour sets, such that the two sets are gender-balanced, and share no speaker; we do a Viterbi time-alignment of the two sets, using the initial 60 hour models. The two time-aligned 30 hour data sets are then used to train two sets of independent acoustic models (although, to make the signal-processing consistent, they do share the parent models’ IMELDA transform).

In the experiments presented here, we fabricate data using one of the 30 hour models, and recognize with the other model. For comparison, we also do two cheating experiments, recognizing with the same 30 hour model that generated the data, as well as the parent 60 hour model [4].

The vocabulary is constructed by taking all the words in the allowable Callhome and Switchboard training sets; there are about 28000 distinct words in this three million word training set, of which 3500 are given more than one pronunciation. All alternate pronunciations for a word are considered equally probable by the recognizer.

The language model is constructed with all the bigrams and unigrams in the Callhome and Switchboard training sets, applying absolute discounting.

## 3. EXPERIMENTS

We present two series of experiments: comparing recognition of real and simulated data, and simpleminded attempts 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 generate data in two different ways: first, using the word transcriptions for the test conversations along with our recognition dictionary (simulating from dictionary, as above), and second, using the phonetic transcriptions for the conversation

(simulating from phonetic transcription). In all cases, we use the first of the 30 hour acoustic models to generate the data from the triphone models. We see in Table 1 that for real data, the two 30 hour trained models produce more or less equivalent word error rates, while the 60 hour trained models are about 2 percentage points better. This is a typical result; it shows that the two 30 hour sets, while yielding comparable recognition results, contain different, and at least partly complementary, information.

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 transcriptions. AM means acoustic model.

However, recognition of the speech simulated from dictionary gives a different picture. When we recognize with the same acoustic models that we used to generate the data, the error rate drops below 5%. This is as if we have trained models on an infinite amount of data (although from a finite number of speakers), in just the right way. Nothing that the data does is unexpected; the model has seen it all before. When we recognize with models trained on completely disjoint data (AM2), the error rate doubles, but still hovers near 10%. We see that the 30 hours of data that AM2 was trained on is different, in some respects, from AM1’s 30 hours. The 60 hour models have seen AM1’s training data, but are led in a somewhat different direction by AM2’s: there is a partial reconciliation, and the result is an error rate intermediate to the two 30 hour models.

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 its inherent confuseability, most of the errors appear to be due to something else!

So if use our recognition dictionary (which has a rather small number of alternative pronunciations for each word) to choose pronunciations, and generate data from these 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 line of Table 1, the data is fabricated using the 30 hour acoustic model 1, along with the ICSI phonetic transcriptions, without recourse to the pronunciations in the recognition dictionary. Word error rates are much closer to those obtained when recognizing real data, than to data simulated from dictionary. Even recognizing with the same acoustic models that generated the data – 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 simu-

lation 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 of conversational speech, the error rate is nearly as high as for real speech. Put more provocatively, variant and reduced pronunciations in casual speech account for most of the errors made by this recognition system.

One explanation for this effect is that by simulating with realistic pronunciations, we may have rendered our dictionary incomplete, as incomplete as it is for recognizing real data. In fact, the phonetic transcriptions match our dictionary less than half the time, leading us to generate data for strings of phonemes that don't match the pronunciation of any word in the dictionary. Previously, when we generated from the dictionary, all of the phoneme strings matched at least one of the entries in the dictionary.

This problem may be made more complicated, but less severe, by the manner in which we train our models. The acoustic models are trained from alignments, in which each frame of training data is mapped to a phoneme state. The phonemes that we map to are determined by the pronunciations in our dictionary, and we know these pronunciations are woefully incomplete for conversational speech. The trainer will encounter several dozen pronunciations for common words in the training data, and try to align them all to the one or two or three prons for the word in our dictionary. The models are smeared, mongrelized to a certain extent, each one forced to represent data for many phonemes, and not just the phoneme they nominally represent. They do partially compensate for out-of-dictionary pronunciations by using multiple components, but consequently are larger, and not as sharp, than they might otherwise be.

### 3.2. Dictionary Augmentation with Simulated Data

Perhaps we can try to improve the dictionary by adding pronunciations to our recognition dictionary. Note that others [5] have also done this with real data; by and large they have not seen improvements in performance. We augment by taking pronunciations for words from the phonetic transcriptions, and adding them to our dictionary even if they occur only once. To gain a sense of scale, we should note that there are about 4700 tokens in the test data, amounting to 900 distinct words; they have altogether 2100 pronunciations. Only 47% of the tokens are pronounced as in our dictionary. About 650 words are pronounced only one way in the test data; *the* has 36 different pronunciations, according to the transcripts. When we add all of the pronunciations found in the test data to our dictionary, only about a quarter are already in our base dictionary, so we end up adding 1500 new ones.

We call this the “base + test” dictionary in Table 2. 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. In fact, this is another instance of a “perfect” dictionary, as in Table 1: each word in the data has its pronunciation in the dictionary. The difference appears to be confuseability: 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, the most “homonymical” pronunciation is associated with five different words: *sons*, *son's*, *sons'*, *suns*, and *sun's*; no pair of words share more than two pronunciations. By contrast, the “base + test” dictio-

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.

nary 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 confuseable pair is *the* and *to*, which have 7 pronunciations in common.

This method of improving our dictionary by adding the pronunciations that occur in the test data is brazen cheating. Suppose we try not to cheat, and use a different set of phonetically transcribed data from which to gather pronunciations: 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 these 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 augmentation.

The “base + train” entry in Table 2 gives recognition results after we have added the 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 confuseability, 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 confuseability added by the extra pronunciations and inevitable homonyms (*the* now has 55 variant pronunciations, and the phoneme *schwa* is a pronunciation for 35 different words; 79 pronunciations have 5 or more homonyms). For the simulated data, it appears that including the correct pronunciations in the dictionary – even if they are hidden in a haystack of dross prons – can be a win, and still improve recognition.

We can see the effects of confuseability in these results by examining the kinds of errors we are making in Table 3. This data is generated with AM1 along with the phonetic transcriptions, and recognized using AM2. 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 recognized using baseline and augmented dictionaries.

### 3.3. Dictionary Augmentation with 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.

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.

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 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 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 the effect increases insertions more and decreases deletions less. Real data is different, however, 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.

This discrepancy may provide more evidence of the mismatch between real speech and our acoustic models, or, equivalently, the dif-

ference between real and simulated speech. Adding new pronunciations to our recognition dictionary appears to add confuseability, and not much else, to recognition of real speech: the recognizer merely has a new sequence of triphones to consider as a hypothesis. This new sequence has not been seen in training – although, quite likely, acoustic training data includes the word being pronounced in that manner, but assigned to a different pronunciation – and the added pron may not match the speech very well. Simulated speech is different: the new pronunciations actually are a good match for words generated from the corresponding phoneme sequence, and so adding pronunciations may yield some benefit (although they also suffer from the deleterious effects of confuseability).

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 one thirty-hour acoustic model, and recognize with a different 30 hour model. We record, for each word token in the correct transcript, 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 this number is 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 correct transcripts, broken down by whether their pronunciations are in the recognition dictionary.

As might be expected, if a word token is pronounced according to the dictionary, it is more likely to be recognized correctly than a token pronounced in an unexpected way. But the difference between the error rates is smaller for real than for synthetic data. It may be that 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. Having the right pron would be relatively more important for simulated data in this view, since, by construction, data generated from a string of phonemes will be a good match for a dictionary pronunciation consisting of those phonemes. Conversely, words for which there is no matching pronunciation would have poor performance, since they do not match well with any of the prons that are in the dictionary.

## 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, word error percentage rates rise into

the 40's; when we attempt to augment the dictionary pronunciations, we see a decrease in the error rate, so long as "enough" correct prons (the ones that occur in the test set) are included. This remains true even when many pronunciations which are not used in the test set are added. Real data, on the other hand, always gets worse recognition results when the dictionary is augmented in this way. We believe this discrepancy is due to a mismatch between real speech and the models we build from them. At least part of this mismatch is due to the extremely varied pronunciations found in conversational speech, and the way which we train our models.

## References

1. Robert Roth, Larry Gillick, Jeremy Orloff, Francesco Scat-tone, Gail Gao, Steven Wegmann, and Janet Baker, "Dragon Systems' 1994 Large Vocabulary Continuous Speech Recognizer", Proceedings of the Spoken Language Systems Technology Workshop, January 22-25, 1995, pp. 116-120
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 Modelling for Conversational Speech Recognition: A Status Report from WS97," IEEE ASRU Workshop, Santa Barbara, December 1997.