Improving Conversational Forced Alignment with Lexicon Expansion

Christopher Liu and **Stephanie Mallard** and **Ryan Silva**
CS 224S, Stanford University
{cwtliu, smallard, rdsilva}@stanford.edu

Abstract

Forced alignment has applications in areas such as accent coaching, speaker identification, and linguistics research where accurate phone-level alignments of speech are incredibly valuable. However, forced aligners for conversational speech are currently not accurate enough for many of these uses. To address this issue, we train and tune acoustic models on an enriched lexicon by applying blanket rules chosen to model informal speech. We hypothesize that by expanding the original lexicon to include varying pronunciations before training and tuning, the acoustic model will be better adapted for conversational speech. We evaluated our approaches with two main metrics: phone labeling and phone boundary accuracy. After experimentation with monophone, triphone, and neural network acoustic models and various combinations of lexicon expansion rules, we achieved 10.8% improvement in phone boundary accuracy over a baseline monophone model with a standard lexicon.

1 Introduction

Identifying phone boundaries in speech is key in tasks for correcting speech, such as giving automatic feedback about accented or misspoken phones and words. Moreover, a critical part of contemporary sociolinguistics research is identifying phones and their boundaries in any given dataset. However, doing this by hand can be a tedious and time-consuming process. Therefore, forced alignment software is a common tool used in order to investigate large amounts of speech data. Unfortunately, standard software such as FAVE-align (Rosenfelder et al., 2011) does not produce accurate enough phonetic alignments for serious research questions involving phone duration, dynamics, or near-boundary measurements in spontaneous speech, and presents a serious trade-off between accuracy and efficiency (Bailey, 2016). The current work-around is to manually correct forced alignments and phonetic transcriptions produced by tools such as FAVE-align, which rather defeats the purpose of using the tool in the first place.

In this paper, we address the challenge of generating accurate forced alignments on spontaneous speech. We explore two methods to accomplish this: first, automatic lexicon enrichment, and second, acoustic model improvement. Because tools such as FAVE-align only use a pre-trained monophone acoustic model trained on the SCOTUS corpus, we will experiment with more sophisticated acoustic models and attempt to improve upon those results. In an attempt to adapt to a spontaneous speech test set, our acoustic models will be trained on Switchboard instead of the SCOTUS corpus. Additionally, we will apply several blanket pronunciation enrichment rules to a standard lexicon. For example, in the case of because, cuz, and cause would be included. Hopefully, these adaptations will assist the forced aligner in adapting to the idiosyncrasies of conversational speech.

2 Related Work

This work is a follow-on to experimentation done by a previous CS224S team, who did work on improving forced alignments using an enriched lexicon (Todd et al., 2014). Forced aligners take in speech and a lexical transcript, and combine a lexicon with an acoustic model to identify a sequence of phones that best describe actual pronunciations, and assign the phones boundaries placed as close

as possible to where an expert linguist might place them manually. This team generated phonetic rules (g-dropping, vowel destressing) to provide an expanded lexicon with additional pronunciation variants. However, they found that this did not improve the accuracy of the phone sequences identified by the forced aligner. Their acoustic model was trained such that phone representations were flexible, which increases the number of lexical alternatives, which resulted in more fitting to noise. They recommended that combining a more highly constrained acoustic model with a flexible lexicon could produce better results.

In regards to improving the acoustic model, both Deep Neural Networks (DNNs) and monophone or triphone Gaussian Mixture Models (GMMs) can be used as acoustic models in conjunction with a Hidden Markov Model (HMM) for decoding raw audio features into a phonetic transcription and producing a forced alignment. However, DNNs have been shown to outperform GMMs as acoustic models when applied to the forced alignment task (Abuzaid, 2013).

Rule-based lexicon enrichment is another technique that has been examined and proven to boost the accuracy of an automatic forced aligner (Milne, 2016). However, this research used a corpus of well-articulated Dutch and did not mention how forced-alignment systems perform on conversational speech. The study also explored how enriching the lexicon during training improves performance in a GMM-HMM system, and concluded that providing a rule-generated lexicon that offered more variants for pronunciations improved forced-alignment quality for all types of GMM-HMM systems.

3 Approach

In general, our work investigates two main methods of forced alignment quality improvement: enhancing the acoustic model (with tuning and model sophistication) and lexicon enrichment. Based on previous work, both have been proven to improve forced alignment in isolation, but have not been tested in conjunction with each other. The main trade-offs we contended with include: over-training the acoustic model and fitting it too closely to the training set, and over-expanding the lexicon to such an extent that the decoding process becomes confused and is no longer able to produce adequate phonetic transcriptions or alignments. Figure 1 depicts our model training pipeline, and the dashed lines indicate the steps we altered to obtain optimal results.

3.1 Datasets and Tools

In order to better conform to conversational speech, our acoustic models trained on the Switchboard Release-2 corpus (Godfrey and Holliman, 1993). Our validation set for use during training was taken from the 2000 NIST Speaker Recognition Evaluation (Eval 2000) corpus (Przybocki and Martin, 2001). The ICSI Switchboard Transcription Project (Greenberg, 1996), a subset of the original Switchboard corpus with manual phonetic alignments, is our test set. To avoid training on our test set, we did remove the recordings used in the ICSI dataset from our Switchboard training set.

The Kaldi speech recognition toolkit (Povey et al., 2011) streamlined the process of training various acoustic models, tuning hyperparameters, and producing forced alignment.

3.2 Acoustic Models

Our experiments included monophone, triphone, and DNN acoustic models. Each one begins training with an alignment generated by the previous model - the monophone begins with an equal spacing alignment, the triphone begins with the result of the monophone model, and the DNN starts with the best results from the triphone model. Consequently, parameter tuning will also be done in that order - monophone, triphone, then DNN in order to maximize final performance. All of these models take Mel Frequency Cepstral Coefficient (MFCC) features as input, and output probabilities of states within a lexicon-generated HMM.

3.3 Decoding

Decoding is performed by the Kaldi toolkit, with the standard Viterbi beam search algorithm for speech recognition. This algorithm takes the acoustic models described in the section above and an HMM generated from a pre-determined lexicon and decides which path, or phonetic transcription, is the most appropriate for the given raw input MFCC features.

3.4 Lexicon Enrichment

Lexicons were generated using Kaldi scripts on the transcripts in the Switchboard corpus. Each

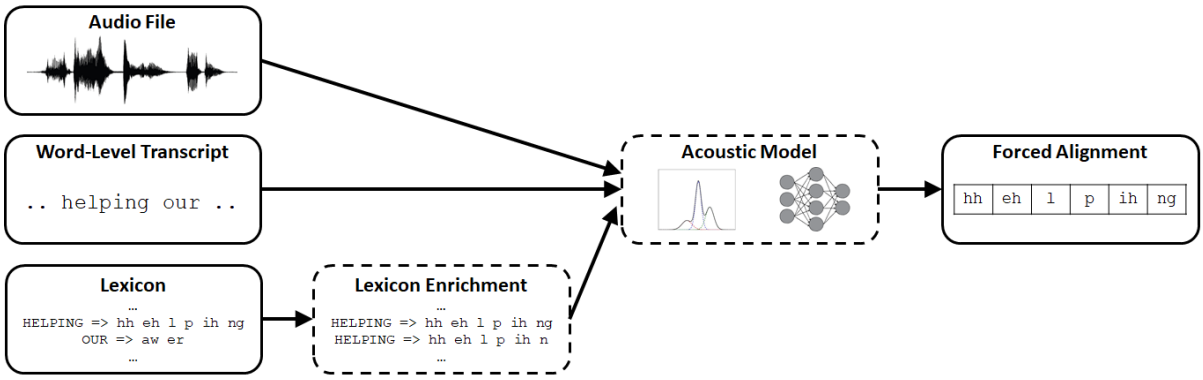

Figure 1: Model Training Pipeline

lexicon contains word entries and its corresponding phone representation, according to lists of phones, silence phones, non-word representations (such as noise and laughter), and out-of-vocabulary placeholders.

After the lexicon was generated, we expanded the lexicon further using a Python script according to various rule combinations. Our reduction rule set is based on pronunciation variations that often occur in conversational speech (Todd et al., 2014).

Note that the original Switchboard lexicon produced at Mississippi State differs from the CMU Dictionary pronunciations. We experiment with both the original lexicon and adding the CMU pronunciations before enrichment.

The rules used for this experiment are listed in Table 1. Our goal was to use varying degrees of enrichment on the lexicon to find the rule combination that allows the Viterbi algorithm to choose the best-fitting pronunciation during forced alignment. Each run was evaluated using a tuned monophone GMM acoustic model. A specific rule set was then chosen for training at the triphone and neural network level.

4 Experiments

Preliminary baseline runs were completed using a monophone Gaussian Mixture Model (GMM), trained on 30k (12hr) of the shortest utterances in Switchboard. We performed hyperparameter tuning using the standard lexicon. We chose the following reference hyperparameters to use for later runs: num_iters=45; max_iter_inc=30; totgauss=1200; boost_silence=1.2; realign_iters =1:1:10,11:2:20,21:3:45.

4.1 Baseline

Our complete baseline run trained first on 30k of the shortest utterances, using a GMM monophone, using standard Kaldi hyperparameters and an unenriched lexicon. A training alignment was computed using 100k utterances with no duplicates for use as a reference alignment to train a baseline triphone GMM model. Both of these models took MFCC features as input.

Afterwards, the baseline neural network was trained on the entire Switchboard corpus with alignments from the baseline triphone model. All neural networks were trained for four epochs over the entire corpus with MFCC+LDA+MLLT+fMLLR transformed features, which reduce the dimension, normalize, and perform speaker adaptation. Reference neural network parameters are: window of four features on either side of the central frame, decaying learning rate starting at .015, and number of hidden layers equal to either 1 or 5. Our results in Table 5 show that phone accuracy and boundary accuracy improve as we increase the sophistication and number of parameters of the acoustic model.

4.2 Monophone Lexicon Enrichment

For our setup, performance was compared across levels of lexicon enrichment using a standard lexicon based on transcription input and a lexicon with additional entries from the CMU dictionary. Using a script provided by FAVE-align, we performed a dictionary look-up for all words in the input transcription and included additional entries from the CMU pronouncing dictionary.

Table 2 shows the results of lexicon enrichment with these two lexicon setups. Each ID rule set incorporates the preceding ID set before it. The

Rule	example usage
g-dropping	substituting NG with N in word-final sequence IH NG
destress uw, destress iy	reduction of IY to IH, and UW to UH
delete td, delete hh	deleting word-finals T and D, and word-initial HH
vowel deletion (function words)	deleting vowels in function words (the, a, an, etc.)
t-initial alveolar cluster simplification	deleting T that occurs before S, CH, or SH
CMU dictionary look-up	broad expansion according to FAVE-align CMU look-up

Table 1: Pronunciation variants applied to the lexicon prior to acoustic model training.

ID	Expansion Rules	Phone Acc.	% Boundary	Phone Acc. (CMU)	% Boundary (CMU)
1	g-dropping	69.0	55.6	68.6	55.5
2	destress uw, destress iy	68.0	53.9	68.3	55.8
3	delete td, delete hh	68.7	53.3	68.6	55.4
4	vowel deletion, TIACS*	67.8	53.1	67.3	53.9

Table 2: Lexicon expansion testing. % Boundary is the percent of forced alignments within 20ms of the hand-generated alignment. Tests build on previous IDs (e.g. 3 include rules from 1 and 2). TIACS* is t-initial alveolar cluster simplification.

results show that percentage of boundaries within 20 milliseconds of the gold set were generally better using the CMU pronunciations incorporated in the lexicon.

One key insight from these results is that lexicon enrichment enhances both phone and alignment accuracy only up to a certain point- that is, we found that three rules for a monophone model produced the best alignments on our test set. When five or seven rules were applied rather than three, the model performed worse, and we propose that this is a symptom of a limited model rather than due purely to lexical expansion. The increased number of pronunciations may be able to be better captured by a triphone model since new pronunciations should produce important context dependent states.

4.3 Triphone Lexicon Enrichment

Tables 3 and 4 show the same lexicon enrichment experiments performed with triphone models trained on about 110 hours of the switchboard corpus and with significantly more gaussians in its acoustic model. There are significant improvements (6-8%) in the alignment accuracies of the triphone models, with the ID3 lexicon (g-dropping, destress uw, destress iy, delete td, delete hh) now performing the best. However, the phone accuracy of these models on our testing set does not significantly improve, and even decreases with some lexicons. These experiments suggest that context dependency is especially important for forced alignment.

Examining the difference between the two triphone runs models, the boundary accuracy is comparable between Table 3 (30k gaussians) and 4 (70k gaussians) for a given lexicon. However, due to computing resource limitations the models were trained for the same number of iterations, so the model in Table 4 (tri2) may require more iterations to fit its gaussians to the larger number of states. Tuning the number of parameters in the model is explored more in depth with neural network acoustic models.

From this data, we selected the rule combination with a lexicon that incorporated CMU pronunciations, and the following rule reductions (g-dropping, destress uw, destress iy, delete td, delete hh), to train a neural network acoustic model with.

4.4 Neural Network Run

We evaluated performance of the forced aligner across trainings using monophone, triphone, and neural network acoustic models. This run used the rule combinations chosen from monophone and triphone lexicon enrichment.

We trained the following neural networks:

NNet with 1 hidden layer, 10.6M parameters/weights, trained with the original switchboard lexicon.

NNet with 1 hidden layer, trained with ID3 lexicon.

ID	Expansion Rules	Phone Acc. (CMU)	% Boundary (CMU)
1	g-dropping	69	61.9
2	destress uw, destress iy	67.9	62.5
3	delete td, delete hh	68.5	63.4
4	vowel deletion, TIACS	67.5	61.6

Table 3: Triphone model (tri1) lexicon expansion testing. Parameters: 3200 leaves, 30k Gaussians

ID	Expansion Rules	Phone Acc. (CMU)	% Boundary (CMU)
1	g-dropping	68.6	62.3
2	destress uw, destress iy	68.7	62
3	delete td, delete hh	68.4	63.3
4	vowel deletion, TIACS	67.4	62.7

Table 4: Triphone model (tri2) lexicon expansion testing. Parameters: 4000 leaves, 70k Gaussians

NNet with 5 hidden layers, 20.6M parameters/weights, trained with ID3 lexicon.

Table 6 shows the phone and alignment accuracy for the networks trained on ID3. The accuracy of the neural net trained with the original lexicon was 0.39 on validation, and the 1 and 5 hidden layer nets with ID3 lexicon scored 0.40 and 0.44 respectively on the verification set. These networks can be trained for more iterations but still provide interesting results.

The improvement between the 1 hidden layer networks trained with different lexicons shows again that enriching the lexicon is beneficial for alignment accuracy. Moreover, this experiment provides some evidence that increasing the number of layers and parameters, and more specifically the accuracy of models, leads to a greater improvement in alignment quality than phone accuracy. From our experiments with neural networks, a 3% increase in phone accuracy led to a 6% increase in alignment. It seems that while a certain level of phone accuracy is required for good alignment, the two are perhaps not as correlated as one might think, and it is interesting to note that between triphone models for the same lexicon (Tables 3 and 4), phone accuracy decreased while boundary accuracy increased for some of the lexicons.

4.5 Phone and Boundary Accuracy

Our primary evaluation metrics were phonetic transcription accuracy and phone boundary accuracy within 20 ms of a manual alignment. The 20 ms tolerance level was chosen on the precedent set by several works (Todd et al., 2014; Yuan et al., 2013). This window is accepted since exact phoneme boundaries are often inherently ambiguous in conversational speech, as even expert linguists often do not agree on the exact boundaries. Future work might also include an RMS value to better evaluate the quality of alignment.

5 Conclusion

From the onset, we set out to evaluate performance using more sophisticated acoustic models and targeted levels of lexicon enrichment to improve forced alignment quality on conversational speech. We found that combining these two modifications produced the highest alignment accuracy.

5.1 Acoustic Models

Increasing acoustic model sophistication improved both phone and boundary accuracy for the standard lexicon and tuned enriched lexicon runs. In practice, FAVE-align, a common forced aligner that uses a pre-trained monophone acoustic model, is not accurate enough for linguistic research purposes when used on conversational datasets. Our findings show that using more sophisticated acoustic models such as triphone and DNN should be preferred when generating forced alignments on conversational speech.

5.2 Enriched Lexicon Models

Both the standard lexicon and CMU pronunciation lexicon revealed that iteratively increasing the enrichment of the lexicon improved forced alignment quality up to a certain point. This may be explained by the fact that rules affect insertions, substitutions, and deletions in WER differently and thus these rules have different impacts on alignment score. There is evidence for this by looking

Acoustic Model	Phone Accuracy	% Boundary within 20 ms
Monophone	68.3	53.4
Triphone	68.8	61.0
Neural Network(1 hidden layer)	69.3	62.9

Table 5: Tuned acoustic models and standard lexicon

Acoustic Model	Phone Accuracy	% Boundary within 20 ms
Monophone	68.6	55.4
Triphone	68.5	63.4
Neural Network(1 hidden layer)	69.1	63.6
Neural Network(5 hidden layers)	69.4	64.2

Table 6: Tuned parameter and expanded lexicon results using ID 3 with CMU

at error changes due to the IY and UW destressing versus T/D deletion rules. For the destressing rule, UW for UH and IY for IH substitutions increased, which led to phone accuracy decreasing, however the alignment quality of this model still went up. For T/D deletion, T and D error deletion in the reference transcript drastically decreased, and both phone and alignment accuracy increased, with alignment accuracy peaking in the triphone GMM model when this rule was added. This suggests that some rules may be more important for alignment quality, especially those rules that affect the insertion and deletion error quantities. Detailed analysis of the types and quantities of phone substitutions, insertions, deletions induced for each lexicon in tri2 and neural net runs is provided below.

5.2.1 CMU Pronunciations

One pattern emerged when comparing phone error types to the baseline, which was substituting AX for AH in the reference transcript was doubled since CMU pronunciations don't have AX in their vocabulary. This is evidence that picking between the two paths in the HMM became a problem for the triphone model, and that enriching the lexicon makes acoustic modeling a harder task.

5.2.2 g-dropping

When only g-dropping rules were applied, forced aligner quality was improved. This revealed that single-rule targeted lexicon enrichment improved phone and boundary accuracy and complements previous research on the presence of g-dropping in conversational speech (Yuan and Liberman, 2011).

However, when examining changes introduced by the g-drop lexicon, there was no significant change in insertions and deletions, in which g-errors were already very low. It is likely this rule accounts for only little increased performance on our testing set.

5.2.3 IY and UW destress

The number of substitutions involving changing the unstressed IH or UH vowel back to its original stressed version doubles or triples, indicating that the model is more likely to choose the wrong pronunciation after new variation is introduced. One interesting detail though is that while this increase in substitution error negatively affected phone accuracy, the system was better able to tell the boundaries of phones.

5.2.4 T, D, and HH deletion

This rule seems to be very helpful for alignment. T insertions into the reference transcript were highest on list of phone insertions and this type of error is now cut in half. The number of HH insertions in the reference transcript was also halved. This said, deletions of these phones also nearly doubles, but the number of deletion errors is much less. The neural net models were also unable to lower insertions and deletions of T and D with the variants introduced by this rule. This may indicate that more training is needed for the neural nets, but it also might indicate that the better performance in alignment we report is simply a symptom of us better targeting our speech data with the hand picked rules.

5.2.5 Performance of rules t-initial alveolar cluster simplification and vowel deletion

The introduction of these rules in the enriched lexicon caused decreases in phone and boundary accuracy for both standard and CMU-pronunciation

lexicons. This decrease may be due to a number of reasons:

Vowel deletion greatly increases insertion error of many vowel phones, thus this model is probably choosing to delete vowels more than it should. Also interesting to note is that the IY for IH substitution error was greatly increased. This shows that when the rules we are adding compound and add multiple pronunciations to a word, they affect each other. This is evidence that the capability of the model to choose the right pronunciation is being affected when exponential variations are added.

Also, the pronunciations variants may not be representative of the dataset. Conversational English varies due to external factors (such as speaker region), so it is important to select rules that are specific to the speech characteristics of both the training and test evaluation set.

The drop in phone accuracy occurs as stacked rules exponentially increase the number of lexical entries. The increase in pronunciation variation is likely to produce the need for more context dependent HMM states which, without enough data support, can cause a decrease in correct phone identification. Because phone identification accuracy and correct boundary placement are to an extent jointly dependent, decreases in phone will inevitably lower the ceiling of what can be achieved in alignment accuracy.

5.2.6 Analysis of neural net output transcripts

The baseline neural net trained on standard lexicon unsurprisingly exhibits the same error patterns as its counterpart GMM model in insertions, substitutions and deletions, with lower quantities.

For the 5 and 1 hidden layer neural nets trained on ID3 lexicon, all types of errors are generally just slightly less in the 5 layer neural net. Comparing the GMM model to the neural nets, both nnets score lower on substitutions and insertions, and in particular UW for UH substitution is halved in the neural net outputs. However deletion type errors are generally higher in the nnet models. Examining the insertion and deletion of silence, it seems that the silence phone is much less common in the neural net output transcripts, as the number of insertions of silence into the reference transcript drastically dropped and the number of deletions slightly rose.

Overall, the neural net models trained on the enriched lexicon had slightly lower errors than the GMM model across all types of phone errors. This indicates that a better model is able to learn the different variations given enough data and training. The analysis of the neural net phone error types is encouraging news that expanding the lexicon can begin to be counteracted with large models and training data.

5.2.7 Comparison to FAVE-align

FAVE-align is most commonly used to generate forced alignments on speech data, but it works unreliably on conversational speech. FAVE-align utilizes a monophone acoustic model trained on oral arguments heard by the Supreme Court of the United States.

Our best neural network outperformed previous work (Todd et al., 2014) in phone identification accuracy. We trained acoustic models on Switchboard, aiming to directly model the wide variation in phonemes in conversational speech and counteract this with increasingly expressive models that are able to achieve high accuracy. From our results it is unclear which type of training data is best suited for forced alignment on conversational speech. Our baseline monophone results were inferior to FAVE-align's monophone model trained on read speech, probably because Switchboard training results in more loosely defined phone boundaries. It is still an open question if triphone and neural network models can adequately adapt to higher variation in phoneme idiosyncrasies and durations, or whether eliminating this variation is better.

Our best DNN boundary accuracy was slightly better (difference of 0.4%) to the FAVE-align results evaluated on the same test set. We were able to show an improvement of 1.3 percent in phone accuracy and 10.8 percent in boundary accuracy compared to our baseline run. Similar analysis using FAVE-align showed a decrease of 0.7 percent in phone accuracy and 0.1 percent in boundary accuracy after lexicon enrichment.

5.3 Future Work

Although we expected that training on the Switchboard corpus would outperform that of FAVE-align, forced aligners might generally perform better using acoustic models trained on formal speech, such as Wall Street Journal or SCOTUS corpus. Continuing this research using targeted lexicon enrichment and more sophisticated acoustic models, trained on a formal speech corpus,

could produce better results.

We also recommend a mechanism of learning specific lexicon rules, rather than applying rule variants broadly across a dataset. In many cases, there are specific sentence contexts when phone reduction occurs. Being able to pinpoint words that should undergo phone reduction similar to actual conversational speech should further increase forced alignment quality. A similar approach is to assign a probability distribution to each pronunciation variation that is learned from the statistics of the dataset, and to trim pronunciations that are under a threshold likelihood.

Forced alignment is still not yet adequate for deployment in high performance systems or in linguistics research use without post-alignment boundary correction. While we did achieve more than a 10% increase over our baseline and showed that lexical enrichment does add accuracy, hand-picked rules are likely not the optimal solution, and additional methods beyond tailored acoustic modeling and lexicon enrichment may be needed to produce reliable forced alignments.

Acknowledgments

We would like to thank Professor Andrew Maas and Jiwei Li for their advice and resources provided during the course of this research. Additionally, we would like to thank Simon Todd for the contribution of his Python evaluation scripts, ICSI Switchboard Project test dataset, and paper from the last iteration of this class.

References

Firas Abuzaid. 2013. Re-alignment improvements for deep neural networks on speech recognition systems. Unpublished manuscript.

George Bailey. 2016. Automatic detection of sociolinguistic variation using forced alignment. *University of Pennsylvania Working Papers in Linguistics* 22(2):3.

John Godfrey and Edward Holliman. 1993. Switchboard-1 release 2 ldc97s62. Philadelphia: Linguistic Data Consortium. Web download.

Steven Greenberg. 1996. The switchboard transcription project. Berkeley, CA. Web download.

Peter M Milne. 2016. Improving the accuracy of forced alignment through model selection and dictionary restriction. *Journal of Phonetics* .

Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, Jan Silovsky, Georg Stemmer, and Karel Vesely. 2011. The kaldi speech recognition toolkit. In *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. IEEE Signal Processing Society. IEEE Catalog No.: CFP11SRW-USB.

Mark Przybocki and Alvin Martin. 2001. 2000 nist speaker recognition evaluation ldc2001s97. DVD. Philadelphia: Linguistic Data Consortium.

Ingrid Rosenfelder, Joe Fruehwald, Keelan Evanini, and Jiahong Yuan. 2011. Fave (forced alignment and vowel extraction) program suite. Http://fave.ling.upenn.edu.

Simon Todd, Guan Wang, and Jingrui Zhang. 2014. Improving forced alignment in conversational american english. Unpublished manuscript.

Jiahong Yuan and Mark Liberman. 2011. Automatic detection of g-dropping in american english using forced alignment. In *Proceedings of 2011 IEEE Automatic Speech Recognition and Understanding Workshop*. pages 490–493.

Jiahong Yuan, Neville Ryant, Mark Liberman, Andreas Stolcke, Vikramjit Mitra, and Wen Wang. 2013. Automatic phonetic segmentation using boundary models. In *INTERSPEECH*. pages 2306–2310.