Evaluating and Correcting Phoneme Segmentation for Unit Selection Synthesis

John Kominek, Christina Bennett, Alan W. Black

Language Technologies Institute
Carnegie Mellon University
{jkominek, cbennett, awb}@cs.cmu.edu

Abstract

As part of improved support for building unit selection voices, the Festival speech synthesis system now includes two algorithms for automatic labeling of wavefile data. The two methods are based on dynamic time warping and HMM-based acoustic modeling. Our experiments show that DTW is more accurate 70% of the time, but is also more prone to gross labeling errors. HMM modeling exhibits a systematic bias of 15 ms. Combining both methods directs human labelers towards data most likely to be problematic.

1. Introduction

Two crucial elements of a concatenative speech synthesizer are the set of prerecorded wavefiles that serve as the basis for voice generation and the indexing of these into a catalog of units. A *unit* is simply a continuous segment of speech, identified by a file index plus starting and ending times, which may be joined with other units during synthesis. The nature of these units is a key classifier of synthesis systems. Phoneme, syllable, half-phone, and diphone-based catalogs are sensible choices that have been tried. Regardless of the type of units employed (including hybrid composition), the quality of synthesized speech hinges on the accuracy of their labeling.

This is an issue because ensuring the accuracy of unit labels remains a labor-intensive process. Automatic labeling algorithms can provide a set of candidate labels from which to begin, but the defect rate is usually too high to avoid a follow-on stage of hand correction. This limits how quickly voices can be created, or conversely, the amount of recording that can be quality-assured within a given time budget. Further, the correction stage (ideally) requires technicians that are fluent in the target language, familiar with acoustic phonetics, and cognizant of the details of current technology – a combination of skills that are not readily available.

So long as some amount of hand correction is necessary, the ensuing drudgery implores us to minimize expended effort. This has three components. First, devising improved automatic techniques. Second, adding confidence information to better direct the effort. And third, one should know the point of diminishing returns. That is, the threshold beyond which additional effort will have little impact on final voice quality.

In pursuit of these goals this paper investigates the type and frequency of labeling errors likely to afflict a unit selection synthesis system, the rate at which errors can be purged from the catalog, as well as the effects felt in the resultant voice. This work is novel in that it combines two independent and automatic labeling techniques, and compares the results against two sets of hand-corrected labels.

2. Software and datasets

Carnegie Mellon's speech synthesis systems Festival, and Festvox, the associated tools for building voices, have recently undergone a new software release [1]. Newly included is the ability to use CMU's SphinxTrain tool – derived from the acoustic modeling portion of the Sphinx-II recognition system – for performing automatic labeling [2]. This is in addition to an existing module that employs dynamic time warping as its alignment technique. Festvox-DTW is less computationally expensive, but is more restrictive in that to be used, a voice must already exist in the target language — or at least exist in a closely related language.

Besides this restriction and difference in speed, why have two? Judged by label accuracy we have found that one technique is not clearly superior to the other, but that the error characteristics are distinctly different. This finding supports the retention of both, and invites the prospect of a hybrid method that performs better than either alone.

2.1. Speech databases

Two speech databases are used in this study. The first consists of the 'F2b' recordings from the Boston University Radio News Corpus, available from the Linguistic Data Consortium [3]. The recordings are prepared news broadcast stories read by an American female with a standard North American dialect. For this work we've removed some of the poorer quality recordings and split others along prosodic boundaries. The result is 155 wavefiles totaling 53 minutes of speech. Summary statistics are found in Table 1.

The second database consists of 534 short sentences read by an American male, also of standard dialect. The sentences are largely excerpted from several famous children's stories such as Alice in Wonderland and The Jungle Book. "There goes the great Mr. Toad" is a representative sentence. We call this database Kal-Text-4. Because it was deliberately designed for use in speech synthesis, the transcript has been enunciated clearly and consistently, in a deliberately flat style devoid of inflection. The speaker is the same voice talent behind the KAL diphone database that is freely available as part of the Festvox system. This is pertinent because the KAL diphone synthesizer is used in the Festvox-DTW labeling method.

For conciseness, most of the results presented below are for Kal-Text-4.

2.1.1. Hand labeling

In order to have a reference for evaluation, the F2b corpus has been hand labeled by two of us (CB, AWB) starting from

automatically generated DTW labels. Our labels are consistent with our lexicon (CMU-Dict) both in the phoneme set and in the pronunciations [4]. Though the Boston University data is packaged with label files, these are based on a larger TIMIT phoneset and so are not directly comparable to ours.

The Kal-Text-4 corpus is notable for having two sets of hand corrected labels, denoted as 1st-pass and 2nd-pass. These are distinguished by the amount of time devoted to each file. A 1st-pass correction can be performed in about two minutes per file with the assistance of a visual labeling tool. Problems that immediately catch the eye include: unnaturally dense conglomerations of phones, uncharacteristically long durations, non-speech, and sections of silence improperly included as part of a phonemic unit. At this stage most gross errors should be removed. The resulting label accuracy is representative of what one can expect from a database prepared for non-commercial use. Three people contributed to this effort. Our F2b labels are also characterized as 1st-pass.

Building upon such a 1st-pass, 2nd-pass correction is more painstaking. Every phoneme is examined in turn, using the waveform, spectrogram, and RMS power curves as visual aids. Moreover, every word, syllable and phoneme is listened to at least once. Label boundaries are adjusted until contamination from neighboring phones is minimized. One can never escape the effects of co-articulation, but at word boundaries at least, the human ear is sensitive to edge impurities. To help ensure consistency in creating a ground truth reference set, only one of us (JK) engaged in 2nd-pass correction. The amount of attention required is an order of magnitude greater; a 2.5 second utterance takes on average 30 minutes to complete. Out of the full corpus of 534, a subset of 100 utterances has been 2nd-pass corrected.

	Bost-F2b	Kal-Text-4	Kal-Text-4b
Ave. dur.	20.58	2.49	2.48
Tot. dur.	3190	1330	248
No. utt.	155	534	100
No. words	8791	4000	685
No. phones	38955	14905	2742
% phone	100.0	100.0	97.6
% diphone	66.4	61.1	40.8
% triphone	11.8	8.4	2.7

Table 1. Some basic statistics of the speech databases. The rows from top to bottom are: average and total utterance length, number of utterances, total number of labeled phonemes (including silences/breaths), percentage coverage of phonemes, diphones, and triphones. Durations are in seconds.

2.2. Two automatic labeling techniques

Dynamic time warping is an algorithm for deriving a nonlinear mapping between two waveforms so that time events in one waveform (here phone-to-phone boundaries) can be aligned to corresponding events in the other. This technique found early prominence as "template matching" in the formative years of speech recognition [5]. Our use is as follows. Given an utterance's text, the Festival front-end translates this into a sequence of phonemes, occasionally inserting pauses at presumed prosodic breaks. Based on this, the 'KAL' American male diphone voice is used to generate synthetic wavefiles. These are time aligned against the corresponding targets. The known phone boundaries from the synthetic wavefiles are then taken as indicating the positions of boundaries in the target wavefile.

In contrast, Festival's SphinxTrain borrows a more contemporary technique from the arsenal of speech recognition, that of acoustic modeling. Speaker dependent, semi-continuous HMM-based acoustic models were built separately from the two datasets. Each HMM comprises 5 states, with single skips permitted. The emission properties of each state are modeled by a mixture of 8 Gaussians, representing feature streams of 13 mel-cepstrum values, augmented by delta and double-delta differences. Context dependent triphone models are state-tied on the basis of acoustic questions, up to a limit of 6000 shared models [6]. This is consistent with the default parameters of Sphinx-II. We adhered to the default parameter settings in order to yield experimental results of interest to typical users. To label database wavefiles, the resulting acoustic model is then used to perform a forced alignment operation. For consistency, the same phoneme transcription is used in both Festvox-DTW and SphinxTrain-based alignment.

Both methods have been set up under favorable conditions. With SphinxTrain, the database to be labeled is the same as is used for acoustic training. For Kal-Text-4, the voice data being labeled is that of the same talent behind the diphone synthesizer employed by Festvox-DTW. The utterances can even be considered spoken in a diphone-synthesizer-like manner. In both cases the results should thus be good indicators of best-case performance.

2.3. Error categories

When referring to labeling errors it is helpful to categorize the ways in which things can go wrong. There are numerous possibilities.

Local alignment errors of boundaries. This is the most common situation. The sequence of phonemes is correct but the endpoints overlap adjacent units or shrink within the correct one. The severity of the error depends both on the degree of misalignment and the particular acoustic units that overlap.

Gross boundary errors. In this case boundaries are so far off that the segment becomes mislabeled as something else. Errors of this type often trigger a cascade of bad labels.

Gross durational errors. In this related problem the duration of phonemes are significantly too short or too long.

Noise injection mistaken as speech. Without having specially trained noise models, extraneous sound such as lip smacks and breaths will be treated as speech.

Substitution errors. This amounts to a deviation of spoken realization from dictionary entries. Use of a lexicon-based front-end opens the door to pronunciation discrepancies. For example the word "to" when spoken may undergo reduction to 't ax' from the dictionary form of 't uh'. Conversely, the front-end may erroneously predict the conversational form of 't ax' when the word was in fact pronounced as 't uh'.

Transcription errors due to OOV words. In the F2b database one frequently hears the tagline "for W.B.U.R." Without an explicit lexicon entry, the phonemic transcript does not sound as "for double you bee you are" but as "wuh burr". Naturally, the frequency of out-of-vocabulary events varies with the language domain. Transcripts derived from news stories are prone to vocabulary misses. This is less of a problem in Kal-Text-4.

3. Measurements

Figure 1 displays a histogram of label timing errors for the Kal-Text-4b corpus. The errors are measured against the subset of 2nd-pass hand corrected labels. A total of 1981 units are compared. This excludes the silence segments that bracket the beginning and ending of an utterance.

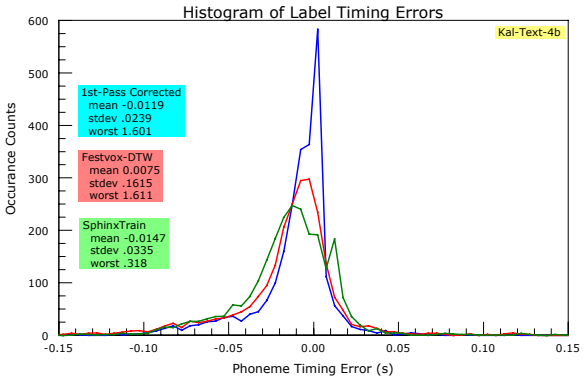

Figure 1. Kal-Text-4 timing errors. Despite the appearance of asymptotic tails, a significant number of events fall outside the graph bounds. The histogram bin size is 5 milliseconds.

Most striking in the plot above is the leftward shift of the SphinxTrain curve. SphinxTrain reveals a habitual tendency to predict the location of phone boundaries too early, by an average of 15 ms. This equates to 1 and 1/2 pitch periods, as the speaker has a fundament frequency close to 100 Hz. Timing errors less than one pitch period are generally okay. Errors larger than four pitch periods are almost always bad.

We attribute this systematic bias to the chain of algorithms and objective functions that control forced alignment. The aligner seeks to maximize the probability of frame sequences according to the built acoustic models, and the acoustic models are Baum-Welsh optimized to minimize word recognition error rates. Segmentation accuracy is not directly a part of this equation. HMM-based models have proved to be quite "inventive" in the solutions they discover. In these experiments they demonstrated an eagerness for the first state to overlap into (what humans would consider) the previous phoneme.

Kal-Text-4	1st-Pass	Festvox-DTW	SphinxTrain
Mode (s)	0.000	-0.0025	-0.0150
Mean (s)	-0.0119	0.0075	-0.0147
Std. Dev.	0.0239	0.1615	0.0335
Max error	1.601	1.611	0.318
pause del	2	9	33
pause ins	4	32	22
sym del	1	1	1
sym ins	5	5	5
sym sub	13	16	15

Table 2. Error statistics of one hand-corrected and two automatically generated labels as measured against the 2nd-pass reference set. Bottom five rows: pauses may be incorrectly deleted or inserted. A non-pause symbol may be deleted, inserted, or incorrectly substituted for another.

Though the half-height widths of the curves in Figure 1 suggest that SphinxTrain has the largest standard deviation, this is not the case (Table 2). Festvox-DTW is more accurate than SphinxTrain in the majority of cases, but also suffers from a greater proportion of outliers.

Error Time Range (s)	1st-Pass Corrected	Festvox-DTW	SphinxTrain
-0.64, -0.32	0	7	0
-0.32, -0.16	3	25	6
-0.16, -0.08	57	112	74
-0.08, -0.04	201	245	302
-0.04, -0.02	267	376	513
-0.02, -0.01	434	480	489
-0.01, -.005	346	307	227
-0.005, 0.0	556	294	187
0.0, 0.005	376	213	190
0.005, 0.01	104	118	127
0.01, 0.02	86	108	251
0.02, 0.04	38	69	74
0.04, 0.08	20	31	23
0.08, 0.16	3	18	23
0.16, 0.32	0	15	7
0.32, 0.64	0	28	0
0.64, 1.28	0	40	0
1.28, 2.56	2	7	0

Table 3. Histogram of label errors grouped into logarithmic bins (linear at the center). Labels of exactly zero error are evenly split between the central pair of rows, shaded green. Rows shaded orange lie outside the bounds of Figure 1.

As expected, the 1st-pass labels exhibit the smallest average error and tightest variance. Nevertheless, since they derive from the tweaking of Festvox-DTW labels, this heritage exerts a bias on the results. The corresponding curves of Figure 1 give an indication of this effect. Notably, two of the most erroneous labels (bottom row of Table 3) eluded capture. The phone-to-phone identities of these pair of bad labels are listed in the top two rows of Table 4.

The worst eight phone-phone transitions for Festvox-DTW are all of the form X-pau, as seen in the second column of Table 4. These often occur at the end of an utterance, but the most dramatic errors are due to internal pause ins/del errors. A misplaced pause can cause DTW to get into serious trouble, from which it never fully recovers. Once removed, the next dominant group involves the voiceless stops p and t (1st-pass column). SphinxTrain's worst offenders are more diverse.

	1st-Pass		Festvox-DTW		SphinxTrain	
1	th-pau	.287	v-pau	.945	zh-ax	.229
2	iy-pau	.140	f-pau	.823	w-ax	.220
3	p-w	.091	th-pau	.756	r-er	.188
4	m-t	.090	t-pau	.687	hh-ao	.113
5	ax-ch	.089	ey-pau	.681	ih-hh	.107
6	ay-p	.087	k-pau	.513	ow-l	.104
7	uw-p	.085	iy-pau	.418	jh-d	.103
8	ey-p	.082	d-pau	.346	ay-p	.097

Table 4. Worst eight phone-phone transitions for each label set ranked by average timing error. Times are in seconds.

4. Error prediction and correction

A linear regression fit between Festvox-DTW and SphinxTrain finds that the pairwise errors are not strongly correlated. If the correlation were strong, having a second set of labels would provide little additional information. As is, the two may fruitfully be combined to identify labels that are likely candidates for correction.

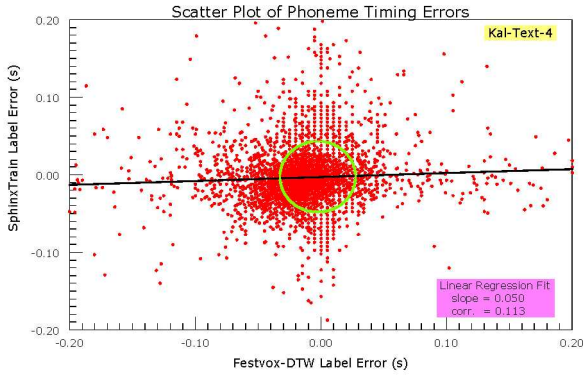

Figure 2. The straight line is a linear regression fit. The green circle indicates the zone of mutual low error. These values are relative to the 1st-Pass label set, chosen for the large number of data points. The coefficient of correlation is 0.114.

We've implemented and compared two straightforward approaches for detecting suspect labels. In one, the duration of labels is used as an indicator of error. Specifically, labels are predicted as bad if their duration is greater than 2 standard deviations from the mean, where the mean is on a per-phoneme class basis. This applies to each dataset separately.

In a second approach the differences between the two label sets is used to compute the average discrepancy per utterance. The human reviewer then examines utterances in ranked order, from worst to best. This is how the 2nd-pass reference was produced. Once an utterance is loaded for review, all phones in that file are corrected, not just the egregious few.

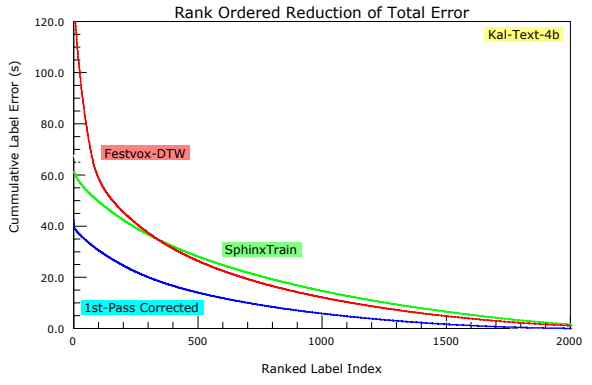

Figure 3. Optimal rate-reduction curves for the label sets. Festvox-DTW has a faster initial rate of reduction than SphinxTrain, due to the larger portion of gross errors. The crossover point occurs after 340 labels are corrected. An 80% error reduction is reached after 798, 519, and 1129 corrections to 1st-pass, Festvox-DTW, and SphinxTrain data respectively. In 70% of cases, DTW errors are smaller than SphinxTrain's.

5. Voice evaluation

To evaluate impact of labeling accuracy on synthesis, we have built eleven voices out of our data. These derive from: 1-4) the four label sets; 5) SphinxTrain, with 15 ms added globally; 6-8) 2nd-pass corrections applied to the other three; and 9-11) using 2nd-pass information to *discard* good labels, keeping a preponderance of bad ones. With this last group, pushing the voice in a negative direction serves to accentuate defects hidden in the larger catalog.

6. Conclusions

The key aim of this work has been to gain insight into the nature of phoneme labeling errors, thereby paving the way for improved algorithms. We've also sought to collect practical advice for voice developers. As guidance, we can offer these observations.

- Festvox-DTW is more accurate than SphinxTrain, when well behaved, but is more prone to gross errors.
- In particular, spurious pauses are the bane of DTW. It also has trouble with the utterance tail.
- The Festival front-end incorrectly adds (or fails to add) pauses. It also introduces substitution errors. Be alert.
- Automatic techniques consistently predict early. This is especially true of SphinxTrain.
- Consider adding 15 ms to every SphinxTrain label.
- Taken together, the two automatic techniques can help locate suspect labels. When lacking both sets, durational information can be used to find probable mistakes.
- Providing two sets of labels to a human reviewer enables better-informed decisions.

7. Acknowledgements

We thank Cepstral LLC for providing access to the Kal-Text-4 database. This work was funded in part by NSF grant "ITR/CIS Evaluation and Personalization of Synthetic Voices." The opinions expressed in this paper do not necessarily reflect those of NSF.

8. References

[1] Black, A. W., Lenzo, K., *Building Voices in the Festival Speech Synthesis System*, http://www.festvox.org/bsv.

[2] CMU, *SphinxTrain: Building Acoustic Models for CMU Sphinx*, http://www.speech.cs.cmu.edu/SphinxTrain.

[3] Ostendorf, M., Price, P. Shattuck-Hufnagel, S., *The Boston University Radio News Corpus*, ECS-95-0001. Dept. ECSE, Boston University, 1995.

[4] CMU, *Carnegie Mellon Pronunciation Dictionary*, http://www.speech.cs.cmu.edu/cgi-bin/cmudict.

[5] Ney, H., *The Use of a One-State Dynamic Programming Algorithm for Connected Word Recognition*, Transactions on Acoustics, Speech, and Signal Processing, Apr. 1984, pp. 263-271.

[6] Singh, R., Raj, B., Stern, R., *Automatic clustering and generation of contextual questions for tied states in hidden Markov models*, ICASSP 1999.