A Family-of-Models Approach to HMM-based Segmentation for Unit Selection Speech Synthesis

John Kominek, Alan W Black

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

Abstract

For segmenting a speech database, using a family of acoustic models provides multiple estimates of each boundary point. This is more robust than a single estimate because by taking consensus values, large labeling errors are less prevalent in the synthesis catalog, which improves the resulting voice. This paper describes HMM-based segmentation in which up to 500 related models are applied to each wavefile. In a listening test of twelve utterances, human judges preferred the proposed technique over the baseline by a tally of 6 to 2, with 4 ties.

1. Introduction

In a unit selection speech synthesizer, one crucial determiner of voice quality is the accuracy of the underlying units, i.e. the segmentation of recorded speech into phonemes. Elements that are mislabeled or have inaccurate boundary times will taint speech with unnatural prosody, awkward timing, and sometimes mispronounced words. Overcoming this problem traditionally requires laborious hand correction of the unit catalog since automatically derived labels are seldom precise.

In [1] we demonstrated that dynamic time warping (DTW) is a technique capable of high accuracy, but is prone to large errors when the alignment run awry. This paper concentrates on outlier identification and correction afforded by the alternate approach of HMM-based modeling. Our choice is motivated by the observation – as has emerged from listening tests – that one big mistake has greater impact on an utterance's perceived quality than several smaller ones.

Our underlying mechanism for segmentation is the technique of forced alignment using Gaussian mixture models. But instead of using one acoustic model to generate a single estimate of label boundaries, in our approach we create a “family” or set of related models, and use these to provide multiple estimates. Units with low variance in their estimates can be treated with high-confidence, and vice versa.

Given enough estimates of each unit's boundary points, taking the average (or median) value is a simple and effective way of avoiding drastic inaccuracies. While we lack a large hand-corrected corpus to serve as ground truth, the listening tests presented in Section 3 demonstrate a measurable improvement in voice quality. Section 2 elaborates our approach where N (the number of acoustic models in the family, hence number of label estimates) exceeds 500.

Such a “heavy duty” attack is novel to this work but is not something to recommend for the casual user. Thus, one specific goal is to save others the computational burden that we have undertaken to explore. In the Festvox [2] voice building toolkit – which we support and employ – training acoustic models for automatic labeling is by far the most time consuming step. Our aim, in a future release, is to include prebuilt models (along with appropriate parameter settings) that have been validated to work well in most cases, at least for English. This would reduce voice building times by a factor of ten, not including the savings of hand correction.

2. Segmenting Speech for Synthesis

An inherent problem of segmentation techniques – whether HMM or DTW-based – is that the program is normally applied with only one parameter setting and yields only single point estimates. An advantage of applying a family of acoustic models is that each segment acquires multiple estimates, and that these results can be combined into hypotheses that are more robust. This section introduced our test data, explains the experimental setup and compares results.

2.1. The CMU ARCTIC Speech Databases

“CMU Arctic” is a collection of studio-recorded, single speaker English databases created with the goal of supporting speech synthesis research. An Arctic database is a reading of the Arctic prompt set by a speaker in a specified style of delivery. The prompt set contains about forty thousand phonemes and about 1150 utterances selected from a 168K utterance corpus [3]. The source text is derived from out-of-copyright novels and short stories.

Where possible, audio is recorded at 16-bit 32 kHz along with a simultaneous EGG signal. The audio and EGG recordings are packaged with phonetic labels, pitchmark files, and related data required to deploy a Festival unit selection voice. These releases use the standard scripts present in Festvox 2.0 to create unit selection voices. They serve as a baseline build for comparison.

Currently, four voices have been released, 3 male and 1 female. These are referred to by the names *awb*, *bdl*, *jmk* (all male), and *slt* (female). A fourth male voice, *rms*, is now under preparation. It is the subject of this study, and is the material we are using to benchmark potential improvements.

2.2. Acoustic Training and Forced Alignment

In [1] we compared phone alignment using a dynamic time warping algorithm to that of HMM-based acoustic modeling, in particular, using the SphinxTrain tool [4]. The primary conclusion drawn was that SphinxTrain is not as accurate as DTW, overall, but is also less prone to large mistakes. It is thus the preferred choice for automated builds.

In the default configuration, SphinxTrain uses a 5-state topology for each phoneme with skips allowed between alternates states. This configuration is recommended for clean speech; in contrast, a 3-state topology without skip arcs is more commonly used under noisy conditions. Speech is processed into vectors of mel cepstral coefficients with a fixed step size of 10 ms. The default Hamming window framesize is 25.625 ms, or 410 samples at 16 kHz. (Sometimes 400 samples are used for a window length of exactly 25 ms.) This relatively broad window equals 2-3 pitch periods for the average male voice. This, and the fact that HMM models are not optimized for boundary accuracy, explains why they cannot be expected to have high accuracy. Using narrower window and step sizes may help, but none of the Arctic voices have hand-corrected labels to serve as reference set for detailed verification.

Unlike [5], which directly attempts to improve boundary placement, the concentration of this work lies in outlier reduction. Our approach is to apply multiple independent acoustic models. The results of all the models can then be averaged to yield a combined decision. This accomplishes two things. First, it avoids the weakness of using one particular parameter setting. Second, the variance of a given unit's boundary times can be taken as a confidence measure. This approach is more direct (though more expensive) that that of [6], which does not build explicit alternate models.

In its default configuration, Sphinx builds models containing 6000 tied triphone states, or *senomes* as they have come to be called.

Training data (hours of speech)		Number of Senomes (recommended)	
1-3	10-30	500-1000	5000-5500
4-6	30-60	1000-2500	5500-6000
6-8	60-100	2500-4000	6000-8000
8-10	100+	4000-5000	8000

Table 1. Degree of state-tieing recommended for a given amount of training speech. These are only rough guidelines.

Table 1 shows recommended settings for training acoustic models to be used in a *recognition* task. Employing 6000 senomes for one hour of speech (the size of each Arctic database) is probably excessive, producing models that are highly voice-specific. Reducing the number of senomes to 500 is better for cross-labeling (using models from voice A to label voice B), but too drastic for self-labeling.

We explored the issue of modeling detail empirically by building separate models in size ranging from 100 to 6000 senomes, in steps of 100. Adding three more – of sizes 250, 750, and 1250 – increases the number to 63 acoustic models per voice. To combat the prospect of over-fitting we also used models from awb, bdl, and jmk to label rms wavefiles. And in addition, each model was run in two modes: with context-dependent triphones, and with context-independent uniphones. (The behavior of ci-models does vary with the number of senomes, particularly in the insertion of silence phones.) Thus each each phone boundary has a maximum of 63x4x2=504 estimates on its position.

2.3. Labeling Discrepancies Between Models

Table 2 contains average discrepancies for labeling the rms Arctic database, in milliseconds. The point of reference in this table is 'rms-median' – the label set produced by using rms acoustic models (self-segmentation) and taking the median value of all 63 estimates. Included are comparisons to the 6000 senome model, and to the average value. The results show high consistency. Recalling that the frame step size is 10 ms, mean discrepancies of 1.38 and 1.45 ms are small. Labels generated using other acoustic models gave an average discrepancy that is larger, from 15 to 21 ms.

Target	Source Model				
rms	rms	awb	bdl	jmk	comb.
average	1.45	17.87	16.74	15.11	8.81
median	0	18.38	17.37	15.36	4.60
default	1.38	21.43	19.63	17.41	—
average	3.54	15.54	14.51	14.36	7.61
median	3.47	15.80	14.74	14.59	9.44
default	3.57	15.81	14.87	14.92	—
accent	american	scottish	american	canadian	mixed

Table 2. Average label discrepancy between rms_median and other systems. The top half represent context-dependent models and the lower half context-independent. The combined model contains up to 252 estimates for each label, whereas each of the others has 63 each. The default configuration uses models with 6000 senomes. Times are in milliseconds.

Results from non-rms voices are not as far away as we initially suspected they might be. Estimates from the awb models differ the most, likely due to the different underlying accent, while the jmk models agree mostly closely.

The combined model incorporates results from the four voice families {rms, awb, bdl, jmk}. With an average discrepancy of less than 10 ms (one frame step), between the combined model and the default, one might not expect a perceivable difference in the resulting voices. And yet, our initial listening tests do indicate that combining labels from multiple models does improve on the default build. We hypothesize that this is due to mitigating the effects of bad units in the unit selection catalog.

2.4. Average Label Variances

The combined model provides up to 504 estimates for each label, in 10 ms precision. The qualifier “up to” is necessary. In cross-segmentation sometimes the Viterbi algorithm that is at the heart of forced alignment cannot successfully find a path through the target waveform. This problem is endemic of using acoustic models so finely tuned to a single voice that state emission probabilities of the target become vanishingly small. Such failed alignment is less of a problem when using coarser acoustic models, i.e. with fewer senomes. This is one argument in favor of parameter tuning.

Table 3 shows that when using all the models in combination, 40% of labels have a standard deviation of less than 10 ms; 70% less than 20 ms; and 90% less than 40 ms. A segment with high variance – larger than 40 ms – does not necessarily mean that the average, or median value is inaccurate. It doesn't even mean that the default labels are inaccurate. But it does indicate which segment boundaries are hard to identify, in the sense that the various acoustic models tend to disagree. The label variance can be used as a measure of segmentation confidence.

Stdev	Cumulative	Stdev	Cumulative
5 ms	7.9%	30	82.7
10	40.0	40	89.1
15	59.7	50	92.8
20	70.4	75	97.4
25	72.7	100	99.1

Table 3. Percentage of segments with stdev in label estimates lower than a given time threshold. Values are for the joint multi-voice model.

Figure 1 displays the distribution of segmentation variances for each of the 40,000 units. The peak value is 8 ms, i.e. less than one frame step. Figure 2 provides a close-up, comparing the combined triphone model (N=252) to rms models alone.

The inference drawn is that models from a single voice tend to agree with each other tightly, at least for the samples we have evaluated. Most of the models place the boundary of a given phone in the same frame. A second peak between 4-5 ms is also evident in Figure 2. This corresponds to a set of boundary estimates that are approximately split between adjacent 10 ms frames. Such tight agreement, we believe, is an effect of being over-trained to a particular voice. It should not be taken to mean that average label accuracy is this low. More likely, overall accuracy lies between 15 and 25 ms, as suggested by Table 2.

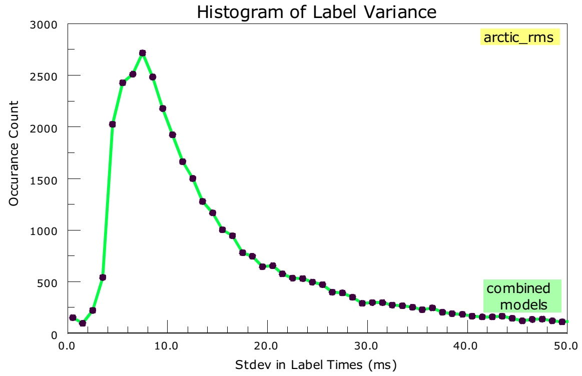

Figure 1. Histogram of label standard deviations by combining the results of each of the 4x63 acoustic models.

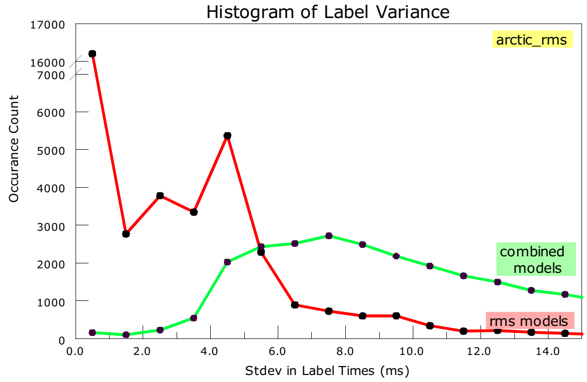

Figure 2. Histograms of label standard deviations, comparing the combined models to self-models (rms). The green curve is the same as that in Figure 1. Note the break in vertical scale at 7000-16000.

2.5. Label Variances in a Waveform

The results of the previous section document overall trends. Here we illustrate with a particular example. Figure 3 graphs label variances for each segment of arctic_a0001, the first prompt of the Arctic recording script. ("Author of the Danger Trail, Philip Steels, etc.")

An interesting pattern is evident from visual inspection. Label uncertainty is greatest at the utterance endpoints, at 1.5s into the waveform, and again at about 2.2s. These correspond to boundaries between silence and speech, either at the ends, or internally. This problem is how significant? If the discrepancy occurred only at the transition from speech to silence (e.g. the sequence /l pau f/ between the words "trail" and "Philip"), it probably would not be severe. However, the graph shows several large error bars occurring in succession, indicating that the problem is not tightly contained.

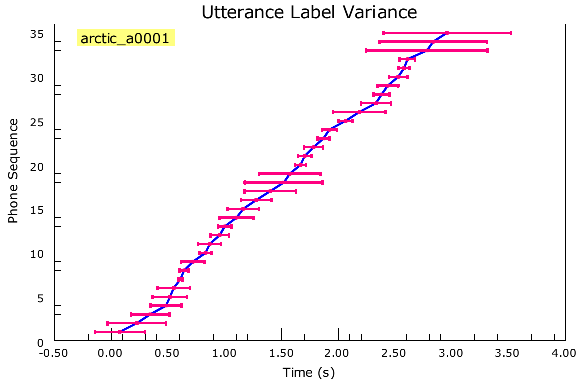

Figure 3. Label variances for the prompt arctic_a0001, using the combined models. Time moves to the right with the dot locating the segment's average end time. For visibility the error bars are exaggerated to show 10x standard deviation. The 35 phones of the utterance are stacked evenly from bottom to top. The phone sequence is: /pau ao th er ah v dh ax d ey n jh er t r ey l, f ih l ax p s t iy l z, eh t s eh t er ax pau/.

2.6. Parameter Tuning

It is useful to ask which single model gives answers closest to that of our family of models. Figure 4 plots the timing differential between single rms models (auto-segmentation), and that of the rms median label values. While not conclusive, this curve suggests lowering the default parameter setting of 6000 senomes to around 3500.

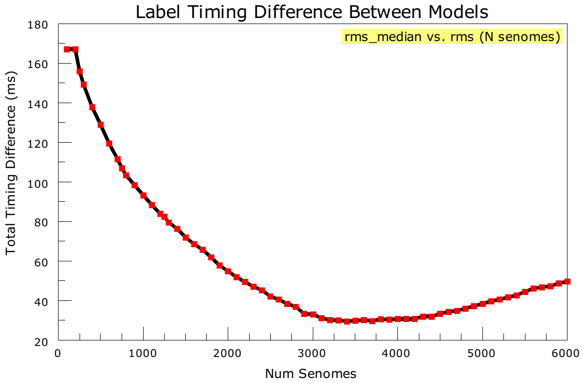

Figure 4. Total label timing differences between singe rms acoustic models and the mean of the entire set (N=63).

3. Listening Tests

Listening tests are crucial to assessing the ultimate impact of a proposed improvement to a synthesis system. Yet, comprehensive listening tests are expensive and time consuming to conduct, and relating the results back to low level system operation is seldom straightforward. The large amount work involved is justified when evaluating major new releases. But for incremental explorations – such as this paper – small scale tests are more suitable.

Our test set consists of twelve utterances drawn from the same source corpus from which the Arctic prompt set was selected. The utterances are single sentences ranging from six to thirteen words in length. They were selected on the basis of having good diphone coverage, with an extra constraint: one of the utterance's diphones must not appear in the Arctic prompt set. (Arctic has good coverage, but it is not exhaustive.) This condition ensures that each test prompt has at least one "difficult bit." There is an exception. One of the test prompts duplicates a training prompt. We put this in to include an example of very high quality synthesis, and to verify that the concatenation algorithm is well behaved. Such an example should be synthesized flawlessly by applying units straight from the source prompt, in sequence. Note that perfect reconstruction is not mathematically guaranteed; it is simply the expected – and desired – result.

For this investigation we synthesized the set of twelve test wavefiles under two conditions. First, using the default procedure for creating the unit catalog (the 'old' set). And second, using the average label times of the combined models (new). We solicited twenty native speakers of American English to render AB preference judgments for each test pair. Testers were allowed to judge a close call as a "tie". The order of wavefiles A and B was randomized but the sequence of pairs remained constant.

After finishing AB comparisons our testers were asked to score each of the new wavefiles on a 5-point subjective scale, with 5 meaning near-perfect. These results are cast against an earlier pilot study in which ten listeners scored the old set of wavefiles. Though such scores are not precisely calibrated between judges, they are helpful for gaining insight into what constitutes synthesis of acceptable quality.

Table 4 summarizes tester responses. Overall, in 6 of the 12 tests, listeners preferred the new results to the old. Two of the cases was considered degraded. Four cases were ties. The absolute scores suggest that the judges in the initial session were generally more lenient than that of the current mix. See for example the scoring of utterances 3, 5, and 12. This is not unexpected. It is far easier for humans to weigh two comparables than it is to be consistent on an absolute scale.

If 1 point is awarded for a 'win', 0 for a 'loss', and half a point for a 'tie', the new wavefiles received 65.8% of possible points from the AB tests. At 95% confidence the performance bounds are [60.3, 70.3] percent.

Utterance	Preference			Score	
	old	tie	new	old (10)	new (20)
1	1	1	18	4.2	4.7
2		8	12	2.3	2.6
3	3	1	16	3.1	3.2
4	7	5	8	2.9	3.0
5	6	7	7	3.7	3.5
6		4	16	3.3	3.6
7	13	2	5	4.6	3.3
8	8	10	2	3.4	3.2
9	4	8	8	3.1	3.1
10		3	17	4.4	4.9
11	2	9	9	3.0	2.8
12	2	14	4	3.0	2.3
Overall	2	4	6	3.42	3.33

Table 4. Results of listening tests. With the 5-point scores a 5 indicates "excellent – almost indistinguishable from real," 1 indicates "poor – intelligible but unpleasant to listen to," while 3 indicates "okay – but with definite defects."

Our intent with the scoring results is to identify wavefiles that serve as reference examples for each of the levels 1 through 5. Once identified, they will be used as calibration points for future, more thorough listening evaluations.

One additional observation, not evident in Table 4, is worth mentioning. In two test cases an incorrect word pronunciation became correct in the new version, i.e. phone substitution errors were fixed. This happened in utterances 1 and 6. The first contains the word "Brokaw" which changed from /b r ow k uh/ to /b r ow k ao/; in 6 the word "suffering" changed from /s ah f eh r ih ng/ (rhymes with 'air') to /s ah f er ih ng/. Substitution errors are an occasional byproduct of labeling errors. Erroneous boundary times sometimes result from a mismatch between the phone identity presumed in the transcript and the spoken realization. Removing outliers, as our technique effectively does, helps subdue this problem.

4. Conclusions

This research demonstrates that applying a family-of-models approach to the problem of segmentation does significantly improve the resulting unit selection voice. In our listening tests 6 of 12 utterances improved while only two worsened. This improvement is due to using averaged segmental boundaries. Our explanation is that this mutes the effect of bad units that otherwise would populate the selection catalog.

Instead of using multiple estimates to revise label boundaries, it is also feasible to remove units altogether. Observe that a well designed unit selection catalog will contain a surfeit of examples in each phoneme category. Up to a point, units suspected of being bad don't have to be corrected; instead, discarding them from the selection catalog is a safe operation. We are currently conducting experiments to discover how this strategy fares.

As for a result that can immediately be put to practice, our experiments suggest reducing SphinxTrain's default configuration of 6000 senomes down to around 3500. Future releases of Festvox can be expected to adopt this change.

5. Acknowledgments

This work was funded in part by the NSF under Grant No. 0219687, "ITR/CIS Evaluation and Personalization of Synthetic Voices." The opinions expressed in this paper do not necessarily reflect those of National Science Foundation.

6. References

[1] Kominek, J., Bennet, C., Black A., *Evaluating and Correcting Phoneme Segmentation for Unit Selection Synthesis*, EuroSpeech 2003, Geneva, pp. 313-16.

[2] Black, A., Lenzo, K., "Building voices in the Festival speech synthesis system," 2000. www.festvox.org/bsv.

[3] Kominek, J., and Black, A., *The CMU ARCTIC databases for speech synthesis,* Technical Report CMU-LTI-03-177, Language Technologies Institute, Carnegie Mellon University, 2003. www.festvox.org/cmu_arctic

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

[5] Matousek, J., Tikelka, D., Psutka, J, *Automatic Segmentation for Czech Concatenative Speech Synthesis Using Statistical Approach with Boundary-Specific Correction*, EuroSpeech 2003, Geneva, pp. 301-04.

[6] Nefti, S., Boeffard, O., Moudenc, T., *Confidence Measures for Phonetic Segmenation of Continuous Speech*, EuroSpeech 2003, Geneva, pp. 897-900.