IMPACT OF DURATIONAL OUTLIER REMOVAL FROM UNIT SELECTION CATALOGS

John Kominek, Alan W Black
{jkominek,awb}@cs.cmu.edu

Language Technologies Institute
Carnegie Mellon University, USA

ABSTRACT

Outlier removal is a straightforward technique for improving the quality of unit selection catalogs without hand correction. This paper investigates the use of phone durations as a criteria for removing bad units. Scoring conditioned on linguistic context demonstrably better than statistics based on phone class alone. The impact of voice modification is evaluated with a 444K utterance test corpus.

1. INTRODUCTION

Unit selection synthesizers are highly sensitive to the accuracy of labeling. Bad labels will adversely affect the quality of synthesis in a number of ways. The phone label itself can be incorrect, potentially causing the wrong word to be said, or said with an undesired accent. Or the label boundaries can be inaccurate – e.g. spilling over into neighboring segments – thereby distorting the speech with impurities. More subtly, bad labels can misdirect the join algorithm, degrading the choices effectively available for neighboring unit selections.

As part of our efforts to improve speech synthesis, we are developing automatic methods for detecting bad unit labels. Two strategies can be taken. The information can be used to direct a human annotator that is correcting the catalog, as in [1]. This is preferred for developing high quality synthesizers. Alternatively, the units considered suspect can be removed outright from the inventory. This is preferred when it is important to build voices of decent quality quickly.

This paper explores voice improvement by removing units based on durational measures. The motivation is that unusually short or unusually long units are most likely mislabeled, and thus removing such durational outliers will be beneficial. Benefit is not guaranteed – a carefully hand-corrected database will still, of necessity, have outliers. Yet the idea is simple, and remains viable when the speech database has an initial redundancy of units. This is true of the CMU ARCTIC single-speaker speech corpus [2] used in this study.

As yet, Arctic databases do not have hand corrected labels. Lacking a reference for evaluation, we have developed an experimental framework for estimating the impact of voice modification. Our new framework is a novel element of this paper.

2. ASSESSING IMPACT

We propose to improve synthetic voices *post-hoc* through manipulation of unit catalogs, as created by the Festvox voice building tools [3]. What impact will doing this have, and how do we know if it is effective?

The intent is that with most of the bad labels discarded – along with some good ones, inevitably – fewer utterances will be synthesized that are plainly unacceptable. “Unacceptable” can be taken to mean that there are no spurious noises present (such as nose breaths), no missing phonemes (caused by inserting a unit of zero length), and no severe substitution errors. Success can be assessed through listening tests, employing utterances known to be problematic. The design of our test suite is described in section 6.3.

Listening tests are bound to be small scale evaluations (or if not, expensive), leaving unanswered the question of overall effectiveness. That is, after a unit catalog has been pruned, how frequently will these changes have a practical effect? It is possible that the selection of outliers is already so rare that the effect of removing them is negligible. The only way to find out, in practice, is to measure this empirically on a large and representative text corpus. After presenting a test suite in section 6.1, our results follow in section 6.2.

As a consequence of removing extreme units from the catalog, the quality of the voice will shift; first subtly, then dramatically. Our prior expectations are this. Initially, unwanted irregularities should diminish. Then, as more units are pruned from the edges, the voice should become more prosodically constrained, more “average sounding.” Accordingly, it may also become better, in the sense that the generated waveforms will converge towards what is predicted from the durational model. At some point, though, there will be so little material available in the catalog that the voice becomes increasingly faulty. This expected pattern is confirmed

through a series of increasingly modified voices, described in section 7.1.

Summarizing, we have a) a series of altered voices, b) small scale listening tests, and c) large scale analytical measures. Together these results help identify a safe range between too little and too much modification. Ultimately, this leads to a voice building procedure that is less error prone, and hence accessible to non-specialists.

3. EXPERIMENTAL SETUP

Before explaining our procedure for modifying synthetic voices through unit pruning, we first describe the data components that are important to this experiment.

3.1. Data Components

The necessary data components of this experiment consist of four pieces.

1. A **speech database** of studio recordings, along with the prompt list and phonemic transcription. Of the four release Arctic databases, we used bdl_arctic.
2. The **unit catalog**. This is an index file mapping all unit instances into the recorded wavefiles. The exact composition depends on the technique used to label the wavefiles. One can label with one to two techniques, in Festival: dynamic time warping on the cepstral feature files (DTW), or HMM acoustic model forced alignment (SphinxTrain).
3. Unit selection **cluster trees**. Cluster trees organize the units of a catalog into divisions that are smaller than simple phoneme classes, constructed on the basis of linguistic context. The tree is a Classification and Regression Tree (CART), built using the Festival program 'wagon' [4].
4. A **durational model**. Similar to 3, this subdivides phone groups into tree-structured clusters. On the basis of surrounding linguistic context, the purpose of a durational model is to predict the length of each target phone in an utterance.

Voice modification is achieved through manipulation of the cluster trees on the basis of durational information. Because of the important role it plays, it is worth pausing to explain the representation of Festival duration models.

3.2. Duration Modeling

By default, unit selection voices in Festival are accompanied by a durational model that has been trained from the 'f2b' voice of the Boston University Radio News Corpus [5]. This is generally acceptable, but it is better to use something speaker-specific. For the experiments of this paper, it is required.

Below is a portion of the durational model that we built for dbl_arctic. This model is part of the publicly available release. The representation is in the form of a Scheme s-structure. Non-leaf nodes are linguistic questions that can be evaluated on a Festival utterance structure. These questions define the structure of the tree. Leaf nodes are pairs the numbers. The second is the zscore mean of a particular cluster, relative to its phone class. The first number is the cluster's standard deviation.

```
((R:SylStructure.parent.syl_break is 4)
 ((n.name is pau)
  ((name is s)
   ((p.ph_cvox is 0)
    ((0.679911 2.86851))
    ((0.660062 1.65406)))
```

This structure says: if the current phone /s/ is followed by a pause and we are at a large phrase break (val 4), and the previous phone – a consonant – has unknown voicing, then in this context an /s/ has an average zscore of 2.86851, with zscore stdev of 0.67991. Otherwise – if the voicing is known – the zscore mean is 1.65406. For example, the final /s/ of the sentence "Those are the breaks." will be predicted to have a duration corresponding to this zscore.

The granularity of the duration model is configured at training time. Ours has 813 leaf nodes trained on 39166 phones. Forty to fifty units is a typical cluster size.

3.3. Durational Z-Score Statistics

Let a phone class p of size n have sufficiency statistics $\{\mu, \sigma, n\}_p$ for summarizing duration times. A particular unit *x* in p with duration d will have a phone-class zscore

$$Z_p(x) = (d - \mu) / \sigma. \quad (1)$$

This equation does not take in account any of the detailed information present in the durational model, so we call it the context-independent score.

When a CART tree is available, a unit will also belong to a particular cluster p,cl with statistics $\{\mu, \sigma, n\}_{p,cl}$. Instead of directly representing durations, these statistics are transformed onto the phone's zscore scale. Let $Z_p(cl)$ be the mean zscore of any unit *x* of cluster cl. Applying (1), $Z_p(cl) = (\mu_{p,cl} - \mu_p) / \sigma_p$. The predicted duration any unit *x* in cluster cl is computed by inversion: $d = \mu_p + \sigma_p Z_p(cl)$.

From here we have two choices for defining a context-dependent zscore.

$$Z(x) = Z_p(x) - Z_p(cl) \quad (2)$$

$$Z(x) = (Z_p(x) - Z_p(cl)) / \sigma_{p,cl} \quad (3)$$

Equation (2) serves to discount a unit's predicted duration from it's actual duration. Thus if a unit is twice as long as the average unit of that phone type – but it is predicted to be twice as long under the circumstances – then it has a context-aware zscore of zero. In other words, its length is exactly as expected. In addition to this compensation, equation (3) rescales (2) according to the cluster's variance in zscore.

These choices have their own merits and drawbacks. If each cluster had several thousand units then the extra precision of (3) makes is preferred. However, the Arctic databases are relatively small, and so the lower confidence of cluster statistics $\sigma_{p,cl}$ argues in favor of (2). Some data presented in section 4.1 lends support for (2).

3.4. Context-Independent Distributions

Figure 1 plots duration histograms for the phoneme /n/. Notable, in this example, is the discrepancy between different labeling methods. On average, SphinxTrain labeled units are 16 ms longer DTW units.

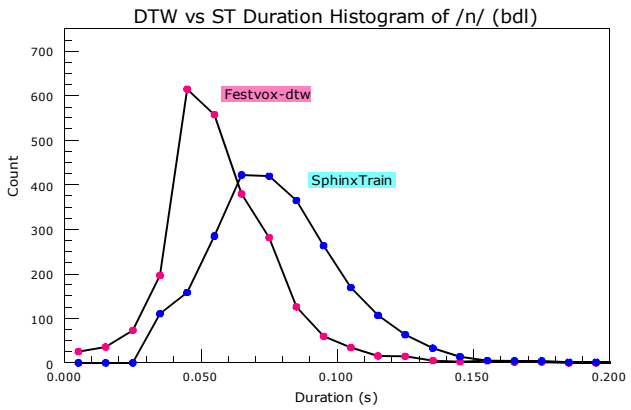

Figure 1. Mean and stdev for these curves, in ms: DTW: (570, 209); ST (733, 255).

3.5. Context-Aware Durational Scores

The following example that illustrates the significance of context-aware durational modeling. Compare the single syllable words 'abe' /ey b/ and 'bay' /b ey/. The phonemes are the same in each, but in reversed order. Consequently, one cannot select the unit /ey/ from 'abe' and expect it to sound natural when used in the word 'bay'. Not only are the acoustics (i.e. formant paths) different, but so too are average durations.

Unit	*Count*	*Mean*	*Stdev*
/b/	627	.0761	.0269
/ey/	619	.1216	.0488
/iy/	1231	.0918	.0443

Table 1. Phone class statistics. Time are in seconds.

Notice that the stdev of the vowels approaches twice that of /b/. Part of this variation is due to the way vowels are extended in length at the end of phrases. The tendency towards extending the voicing of phrase-final vowels is represented in context-aware durational models. Contrast the CART tree predictions for our pair of sample words. Both occurrences of /ey/ are longer than is average for its phone class (zscore > 0). But the word final /ey/ of 'bay' is more than two standard deviations away from the mean, leading to a predicted duration of 225 ms.

word	**ZS**	**time**	**word**	**ZS**	**time**	**word**	**ZS**	**time**
/ey/	0.897	0.165	/b/	0.140	0.080	/b/	0.140	0.080
/b/	0.484	0.089	/ey/	2.123	0.225	/ey/	0.897	0.165
abe		0.255	**bay**		0.305	/b/	0.484	0.089
						babe		0.334

Table 2 Durational model predictions from bdl_arctic. The pattern of 'babe' is that of 'bay' + 'abe'.

The Arctic prompt set contains no examples of the word 'bay', but it does have examples of 'day'. The prompt arctic_b0505 ends in “that first day” while arctic_b0321 end in “the second day.” The durations of these phrase final /ey/ phones is 260 and 230 ms. The predicted duration of /ey/ in 'bay' (identical to that found in 'day') is thus perfectly reasonable; if anything, it is slightly short.

Word	**ZS**	**Time**	**Word**	**ZS**	**Time**
/b/	-0.160	0.072	/b/	0.073	0.078
/ey/	-0.111	0.116	/ey/	2.160	0.227
/b/	-0.478	0.063	/b/	1.346	0.112
/iy/	0.604	0.119	/z/	1.956	0.143
baby		0.370	**babes**		0.560

Table 3. Effect of appending a fourth phone.

In table 3, appending /iy/ to the word 'babe' quickens the pace of delivery, while appending /z/ slows it down. Observe that the final phoneme significantly affects not just one, but the two preceding segments.

What this examination supports is the advantage of using speaker-specific, context-aware durational models for the purpose of outlier identification. Units with a zscore that is two, three, or more standard deviations away from the phone-class average are not necessarily bad on that account alone. They still may have a role to fulfill, in certain contexts.

3.6. Experimental Procedure

The following steps outline how to create a series of pruned voices.

1. Build a voice using the Festvox methodology. This is the “base voice.”
2. With this base voice synthesize all utterances in the Arctic prompt list. Note the identity of all units.
3. Compute mean and stdev values for each phone class. In combination with step 2, compute the context-independent zscore of each unit in the catalog (eq 1).
4. Build a speaker-specific durational model. This offers predicted zscores for unit clusters. In combination with step 2, compute the context-aware zscores of each unit in the catalog (using equations 2 and 3).
5. We now have 3 lists of z-scores, any of which may be used. Establish a pruning threshold, e.g. 5. Find the set of units with absolute zscore greater than this threshold.

6. Build a modified voice by removing the thresholded units from the unit selection cluster tree. Synthesize a set of test utterances from this modified voice.
7. a) Single-pass variation. Repeat steps 5-6 for a series of decreasing thresholds.
   b) Iterative variation. After removing units from the previous step, re-run wagon to construct new cluster trees and a new duration model.

This procedure distinguishes between iterative and single-pass pruning In these experiments we have done single-pass pruning. The iterative approach is more correct, but also more time consuming.

4. COMPARING SCORING CRITERIA

Applying the three scoring criteria will result in different prune lists. A pattern is apparent in Table 4. At a given threshold level the context-aware zscore of (2) reduces the estimate of bad labels claimed by (1), e.g. from 495 to 231. This suggests that a durational model can distinguish between units that are duration extremes, versus those that are inappropriate for the particular context. However, when using (3) vastly more labels are deemed outliers. This is probably an exaggeration.

Z-Score	*Festvox DTW*			*SphinxTrain*		
Threshold	*eq 1*	*eq 2*	*eq 3*	*eq 1*	*eq 2*	*eq 3*
10	1	1	15	2	3	55
8	4	4	37	15	12	102
6	21	15	111	39	28	246
4	154	108	476	185	136	840
3	495	341	1283	426	342	1742
2	1894	1318	3831	1421	1481	4434

Table 4. Number of units above specific zscore threshold.

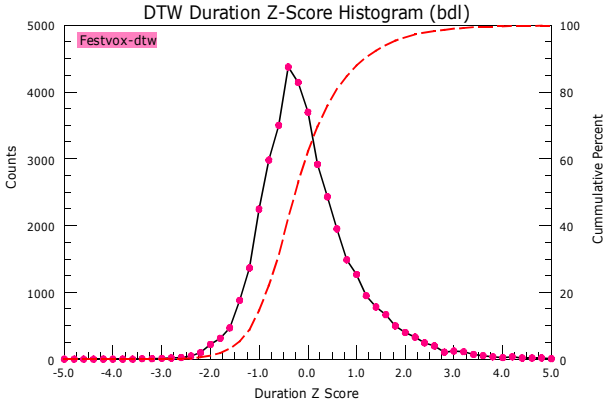

Figure 2. Histogram accompanying the first column of Table 4.

A related comparison involves holding fixed the number of units pruned, measuring mutual overlap. The set intersection curves of Figure 3 show that the criteria of (3) agrees much more closely to that of (2) than of (1).

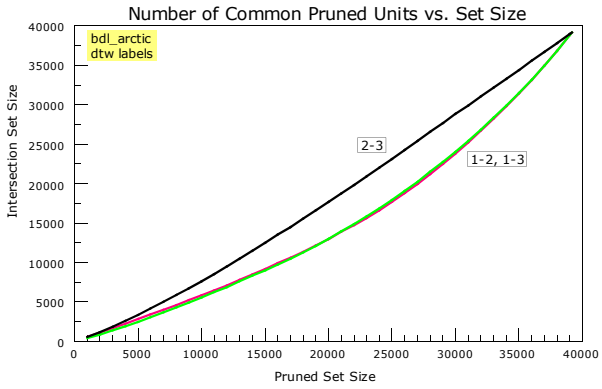

Figure 3. Intersection of 3 pairs of prune lists.

4.1 Scoring Comparison of Worst Outlier

We now illuminate the top row of Table 4 by examining the contents in detail. Table 5 lists the unit id, prompt name, and zscore for the top two durational outliers.

Unit	*Festvox DTW Labels*			*SphinxTrain Labels*		
Rank	*eq 1*	*eq 2*	*eq 3*	*eq 1*	*eq 2*	*eq 3*
1st	l.974	l.974	ax.1812	z.672	z.672	n.1831
	a373	a373	a292	a373	a373	a290
	15.55	14.46	18.12	16.39	15.87	27.13
2nd	n.1752	g.255	l.974	ax.831	ax.831	ax.831
	a329	a387	a373	b153	b153	b153
	8.98	8.95	15.87	13.17	13.50	27.02

Table 5. Top 2 units. Highlighted is the top ranking prompt. Row key, top down: unit id, prompt id, and zscore value.

By consensus, the most problematic recording is prompt a0373 (“Points of view, new ideas, life.”) at the word boundary bewteen *ideas* and *life*. At this position the voice talent put in a lengthy pause. Unfortunately the phonemic transcript does not have a pause marked at this location. The two labeling techniques took opposite tacks: DTW assigned most of the pause to the following /l/, while SphinxTrain assigned it to the preceding /z/. This is reflected in the corresponding zscores.

In the rightmost column of Table 5, the score of unit z.672 places it further down the list, in 45th spot. Examining some of the units higher up reveals that this ranking is erroneous. Observations such as this lead to the conclusions that when cluster sizes are relatively small, the rescaled context-aware zscores of (3) are less reliable than those of equation (2).

5. UNIT PRUNING CASE STUDY

To illustrate the effects of unit removal, here we examine a short example in depth. The test utterance is “Ah, oui.” – /pau aa w iy pau/. Measured by density of outliers, this utterance ranks highly impacted. Seeing it before and after resynthesis illuminates both the benefits and pitfalls of unit pruning.

Item	Score	"ah, oui."			Score	
	pred.	unit	utter.	dur.	z-score	z-s diff
/pau/	0.087	p.1904	a0461	0.560	-0.170	0.257
/aa/	0.697	aa.499	a0461	0.424	7.534	6.837
/w/	1.146	w.533	a0461	0.083	0.907	0.239
/iy/	0.604	iy.891	a0327	0.267	3.468	2.864
/pau/	-0.520	p.2288	a0327	0.178	2.672	3.192
total					14.751	13.389
a0461	"**ah**, we were very close together in that moment."					
a0327	"they were less stooped than **we**, less springy in ..."					

Table 6. Units selected for the test utterance "ah, oui." using the base bdl voice. Highlighted in orange are units with zscore > 3.

With the base voice bdl_arctic_dtw, start by synthesizing the test utterance "ah, oui." Table 6 provides information about the units selected, including: unit id, unit duration, prompt file it is extracted from, the context-independent zscore, the predicted zscore and the difference between the two. Two prompts are used during synthesis: a0461 and a0327. One might expect the utterance to be synthesized entirely from the beginning of a0461, but poor automatic labeling interferes with this continuation.

At a zscore threshold of 3, two of the five phones are durational outliers. But which two these are differs, depending on the criteria used. If simple phone-class statistics are used, the unit iy.891 is considered bad with a score of 3.468. Adjusting the zscore by subtracting the predicted value brings this unit below threshold – but pushes the following unit above (pau.2288).

Listening to the synthesized wavefile does confirm that something is amiss in both of the the final two phones. Specifically, pau.2288 incorrectly extends into the following phone /l/, while iy.891 happens to contain undesired breath noise at this phrase boundary. The first half of the utterance – containing the /aa/ – does not sound bad. It just happens to be unusually long, taken as it is from a single syllable phrase "ah,".

	"ah, oui."				Total	
	/pau/	/aa/	/w/	/iy/	/pau/	z-s diff
none	a0461	a0461	a0461	a0327	a0327	9.940
aa.499	a0143	a0143	a0461	a0327	a0327	4.581
iy.891	a0461	a0461	a0461	a0291	a0446	8.266
both	a0143	a0143	a0461	a0291	a0446	2.906
sphinx	a0461	a0461	a0461	a0461	a0484	6.112
a0143	"**ah**, i had forgotten, he exclaimed."					
a0291	"the w**ee**ks had gone by, and no overt acts ..."					
a0484	"no-sir-ee_."					

Table 7. Units sources after resynthesis, plus synthesis using SphinxTrain labels. Highlighted in orange are units from a0461.

Table 7 charts the effect of removing the unit aa.499 from the catalog, then returning it and removing iy.891, and then removing both. With aa.499 removed the search chooses another unit from the same cluster, this time corresponding to the beginning of file a0143 (with the preceding pause). In the rightmost column we see a drop in zscore error, indicating that the new unit for /aa/ is shorter. As it happens, this replacement doesn't remove the wavefile's principal defect. That occurs only after iy.891 is removed from consideration. Interestingly, the new unit chosen (from a0291) is not from another example of the word 'we', but is the medial vowel of the word 'weeks'. With both bad units are removed the result is an amalgam of the previous two cases.

A couple conclusions emerge. First, removing outliers can avoid poor choices that otherwise would occur. However, any given outlier – even relative to predicted durations – may still be good and doesn't necessarily deserve to be deleted from the unit catalog. Also, this examination underscores the importance of accurate labels; the results of the join algorithm depends acutely on their exact placement.

6. IMPACT ASSESSMENT

Out of all the billions of potential utterances a user is likely to supply, the creator of a voice wants to know which will sound better and which worse. Testing this directly is clearly impractical.

What is possible is an indication of *impact*. Impact is defined as the ratio of utterances that are synthesized differently from the base voice. We estimate impact with a collection of representative text. This is synthesized in the base voice. Then for each modified unit catalog, resynthesize the test suite, noting those utterances now comprised of different units.

A less expensive approximation is to synthesize the test suite once, scanning for units with zscore above predefined threshold levels. Any such utterance is "impacted"; others are assumed unchanged.

6.1 Impact Text Suite

As described in [2] the Arctic prompt set is derived from a larger text corpus of 168K utterances. To this we added the Wall Street Journal (WSJ) and Broadcast News (BN) corpora. Together these cover the intended domain of application of Arctic voices (fictional story reading), and that of a reasonable extension (news story reading). We used the English frontend of Festival to convert the raw text into approximately sentence length utterances.

Corpus	Occurrence Counts			
	Utts	Words	Phonemes	Diphones
arctic prompts	1132	10,045	39,166	38,021
arctic full text	168,443	2,545,156	9,541,969	9,309,645
wall street journal	91,255	2,112,017	9,367,713	9,276,458
broadcast news	184,993	2,851,163	11,071,463	10,886,337
combined	444,824	7,508,336	29,917,236	29,472,440

Table 8. Composition of 444K text suite.

The combined text suite has 444K utterances containing a shade under 30M phonemes. This is a respectable amount, but needs to be placed in context.

Item Types	Corpus Coverage			
	Arctic	*Combined*	*Maximum*	*%*
phones	41	41	41	100
tree clusters	1287	1,287	1,287	100
catalog units	36758	37,601	39,166	96.0
di-phones	1,532	1,592	1,655	96.2
di-clusters	138,630	178,521	1,640,690	10.9
di-units	1,144,670	1,999,807	1,514,069,436	.0013

Table 9. Coverage of 444K combined text corpus.

Table 9 shows coverage for six kinds of items. The base voice has 39166 units of 41 phone classes, organized into 1287 tree clusters. Of these 39K units, 37601 (96%) are used at least once to synthesize the combined test set.

In accordance with the lexicon employed by Festival for American English, there are 1655 possible diphone pairs. 96% of these are encountered in the combined text corpus. Yet, with 1.5 billion unit digrams available in bdl_arctic, we don't even approach exhaustive coverage of all segment joins. Counting clusters offers and intermediate position. To place guarantees on the quality of coverage, we conjecture that there must be at least one good example for each possible di-cluster join. Our test set covers 11% of the 1.6 million possible di-clusters.

6.2 Impact of Unit Pruning

Table 4 shows the number of units that or over-threshold for each of the three criteria. The corresponding measure here is the number of test utterances that are affected due to voice modification. A rule of thumb might be that the impact is substantial when half of the utterances are affected by the voice modification. Using eq. (2) this corresponds to a zscore threshold of about 2.5.

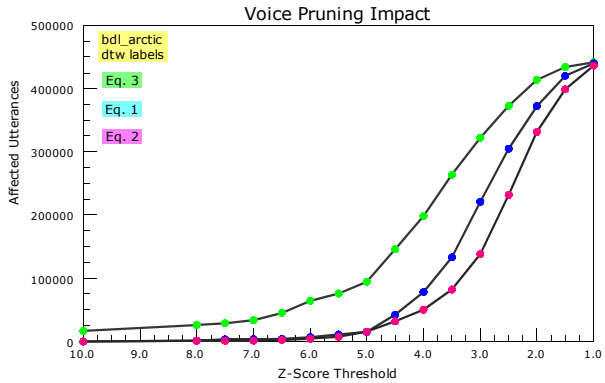

Figure 4. Comparison of impact as a function of z-score.

6.3 Utterance Selection for Listening Tests

Selecting a small yet informative set of utterances for listening test is challenging. However, the approach of the previous section offers assistance. For a given threshold level we know which utterances are altered, and which are not. Even if the test corpus as a whole can be taken to be representative, it makes little sense to test those that won't be affected by voice modification. Excluding these, there still remain an abundance to choose from. We've adopted the following heuristic to select test sentences.

If an utterance u is n phones long and is synthesized with b bad units, then score it with $S(u) = (b/n)\, G(\mu,\sigma)$. G is a Gaussian weighting introduced to favor utterances of a desired length μ. Otherwise, if ranking is based solely on the number of bad units, very long utterances rank highest. Conversely, using the ratio b/n will favor a few very short utterances (such as "ah, oui."). A reasonable parameterization for weighting is $\{\mu,\sigma\} = (24,8)$.

To create a test suite:

1. Rank the utterance set found with (an extreme) zscore threshold of 10. Call this U_{10}. Select a portion of these. Say, ten or twelve.
2. Drop the threshold to 9, yielding the set U_9. Select another ten utterances from the diffset U_9-U_{10}.
3. Continue to some stopping point, e.g. zscore = 2.

The idea behind our sampling method is to organize test utterances into layers. In the first layer, defects found in the synthesized waveform should be plainly obvious – and easy to fix. At lower layers defects will be more subtle, perhaps not attributable to labeling problems at all.

7. CONCLUSIONS

Out research has reached the stage of preparing listening tests for 10-20 users. The results should reveal to what degree the methods introduced here support the automatic construction of high quality voices.

Achieving this objective will enable broader and more varied adoption of speech synthesis technology, our long term goal.

8. ACKNOWLEDGMENTS

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.

9. REFERENCES

[1] J. Kominek, T. Bennett, A. Black, "Evaluating and correcting phoneme segmentation for unit selection synthesis," EuroSpeech 2003.

[2] J. Kominek and A. Black, "The CMU ARCTIC databases for speech synthesis," Tech. Rep. CMU-LTI-03-177, Language Technologies Institute, Carnegie Mellon University, 2003. http://www.festvox.org/cmu_arctic.

[3] A. Black and K. Lenzo, "Building voices in the Festival speech synthesis system," 2000. http://festvox.org/bsv.

[4] A. Black, P. Taylor, R. Caley, "The Festival speech synthesis system," 1998, http://festvox.org/festival

[5] M. Ostendorf, P. Price, S. Shattuck-Hufnagel, "The Boston University Radio News Corpus," Technical Report ECS-95-001, 1996.