CMU Blizzard 2008: Optimally using a large database for unit selection synthesis.

Alan W Black, Christina L. Bennett, John Kominek,
Brian Langner, Kishore Prahallad and Arthur Toth

Language Technologies Institute,
Carnegie Mellon University, Pittsburgh, PA.
{awb,cbennett,jkominek,blangner,skishore,atoth}@cs.cmu.edu

Abstract

This paper describes CMU's entry for the Blizzard Challenge 2008. Our eventual system consisted of a fairly conventional layered cluster based unit selection system using the most predictable subset of the whole UK speech databases. This paper describes the methods we used to find the most reliable subset and the techniques used to optimize the selection. An additional technique that was used was to automatically detect "hard" text and modify the phrasing algorithm accordingly. Although this technique was targeted at SUS utterances it was in place for all utterances. CMU's entry is letter M in the results.

Index Terms: Speech Synthesis, Unit Selection, Database Pruning.

1. Introduction

This paper describes Carnegie Mellon University's entry for the 2008 Blizzard Challenge. The submitted system is a basic unit selection synthesis system; however, it was built using data obtained from a statistical parametric system. At the start of the effort, we were unsure if our core CLUNITS [1] – a standard unit selection synthesizer – or our CLUSTERGEN [2] statistical parametric synthesizer would be best, so we carried out most of our experiments competitively between the two options, only deciding which system to submit as our entry towards the end of the training period.

This year's challenge provided two new speech databases for voice building. The primary one is a UK English database. The second database was Mandarin Chinese. Although basic builds were made, our group did not submit a Chinese system.

This paper describes the database, the selection of the best subset, the three level unit selection system, the weight tuning techniques, and the break predictor system for unusual text.

2. Database Pruning

The core UK English database (Roger) contains 9508 utterances, consisting of around 16.5 hours of speech. From our experience with other varied-style databases, we know that labeling accuracy is very important for both unit selection and statistical parametric synthesis. Our first stage was to find out how well we could label (and model) the different parts of the given databases. We split the data into different sections, News, Arctic, Conversation and Emphasis. Using EHMM [3] we labeled the whole data as well as the different subsets. Using CLUSTERGEN [2], we built statistical parametric synthesizers for different subsets and tested how well we could predict held out test sets. We used Mel-Cepstral Distortion for spectral modeling, RMSE for F0, and Correlation for duration modeling. We quickly noticed that we were not getting good models for the Conversation and Emphasis data, and so dropped those subsets from our builds.

Thus, for our further tests we used only the News (2384 utterances) and Arctic (1132 utterances) subsets, with just over 7.5 hours of speech – less than half of the original data.

We did further experiments ordering the data using its maximum likelihood score from labeling, and checking if high-ranking utterances provided better synthesis than random (or low-ranking) ones. However, this proved not to make any significant difference.

3. Phonemes and Post-lexical rules

The provided utterances already had some form of post-lexical rules applied. As it was not clear what exactly that was, we investigated generating our own phonetic strings (and boundaries). We used the Edinburgh UniLex Lexicon [4], but built our own letter-to-sound rules with it. Again, we used CLUSTERGEN models to measure our objective success in synthesis; we found our own predicted phone streams to be better than the provided ones.

We also noted the almost non-existent post-vocalic R due to the Roger voice's accent, a somewhat classical British English RP. After several experiments to find the best solution, we introduced a new phoneme "rr", which we used explicitly in all post vocalic-R positions. That is, we did not implement R-deletion in our entry, which is a common method to deal with this dialect phenomenon. We re-labeled the data with the "rr" phoneme using EHMM.

4. CLUSTERGEN vs CLUNITS

As we were unsure if we wished to use CLUSTERGEN [2], a statistical parametric synthesizer technique, or CLUNITS [1], we continued to have internal competitions between the approaches to find the better technique.

CLUSTERGEN has since added maximum likelihood parameter generation (MLPG) to its arsenal, but testing of Global Variance did not improve the quality. A second addition was pruning frames from the training data whose acoustic voicing differed from their phonetic voicing. Using standard voicing tools, we labeled each acoustic frame with a voicing flag. We then compared that to the phonetic label and if they differed, we dropped the frame. We are aware this is overly exact, and some unvoiced consonants may be voiced in some context; likewise, some voiced consonants may be validly unvoiced. This technique, which we referred to as the VUV system, made the synthesis substantially smoother and removed some huskiness from the output.

In CLUNITS, we investigated changes of the binding of phone labels, and adding previous, next, and syllable-level tags. It should be noted that this is extending the number of units in the ASR triphone model (context dependent units), rather than lengthening the units. We also re-labeled with EHMM for each of these extensions (cf [5]).

4.1. Internal Listening Tests

We then conducted listening tests for each of the five systems – three CLUNITS voices, and two CLUSTERGEN voices, which we assigned the internal letters A through E.

- A – clunits baseline
- B – clunits with context name binding
- C – clunits with syllable level binding also
- D – clustergen baseline
- E – clustergen with VUV

Aware of the fact that unit selection and statistical parametric synthesizers have dramatically different acoustic qualities, and that we wanted to choose between them, we opted to perform three sessions of AB-tie listening tests. Each session asked a different question.

- Q1. Which is more human-like?
- Q2. Which is smoother?
- Q3. Which do you prefer?

The third question is the normal request of the listener – to weigh all factors in reaching a preference decision. The second question asks the listener to isolate the smoothness quality in the presented samples. This concords with the observation that CLUSTERGEN voices are smooth but not natural sounding. Complementary to this we could have asked "which is spectrally brighter (richer/realistic)?" – keying on the observation that unit selection is capable of being highly natural but is often degraded due to bad joins – but decided that such phrasing was nonstandard and overly specialized. Instead the question asked about human-likeness. Feedback from test subjects indicated that they considered the unit selection voices to be more human-like in general, unless the effect of bad units and/or bad joins was too severe. In which case the smoother CLUSTERGEN waveform was considered more human-like.

Our goal in the listening tests was to quantify and relate these three dimensions, with the task of selecting the best system for submission. For any presentation of wavefile pairs, if the user selects A, then the system that synthesized it accumulated 1 merit point. When the users selects "tie" then A and B split the point. The best system, CLUNITS B scored 65% on all the listening tests combined, followed by CLUSTERGEN E at 55%. B was then developed further to support multilevel unit selection (see Section 5), and submitted as our entry.

While computing a percentage score is straightforward, it has disadvantages. It does not take into account the difficulty of "opponent" that each system faces and the number comparisons against each one. Nor does it provide confidence bounds. These shortcomings are addressed by casting the AB-tie tests as a Bradley-Terry problem.

Sys	Human n=100	Smooth n=100	Prefer n=153	Comb. n=353	LOS (%)	Score (%)
B	109	-4	106	143	93	65
E	-32	81	48	96	92	55
D	-70	90	-43	51	82	47
C	39	-62	-66	21	75	43
A	-47	-105	-44	0	–	39

Table 1. Ratings of five systems according to the 3 questions, and combined. The Combined column has been rescaled so that the worst system E has a rating of 0 and has 90% confidence intervals of ±40. LOS is "likelihood of superiority" over the next best system.

The Bradley-Terry model is a simple and much studied means to describe the probabilities of possible outcomes when individuals are judged against each other in a set of paired comparisons [6]. The central idea is to convert the results of all A versus B comparisons into a scalar rating for each system. The rating represents the system's overall strength, where equal differences in ratings imply an equal difference in outcomes. The probability that A beats B obeys a logistics sigmoid function. In our calculations p(A beats B) = 0.5, p(A ties B) = 0.26 and p(A loses to B) = 0.24 when A is 100 rating points above B. Efficient algorithms exist to maximize the likelihood of the data given the model. The ratings of the five systems are listed and Table 1. Also, the ratings for the two dimensions *Smoothness* versus *Humanness* are plotted in Figure 1.

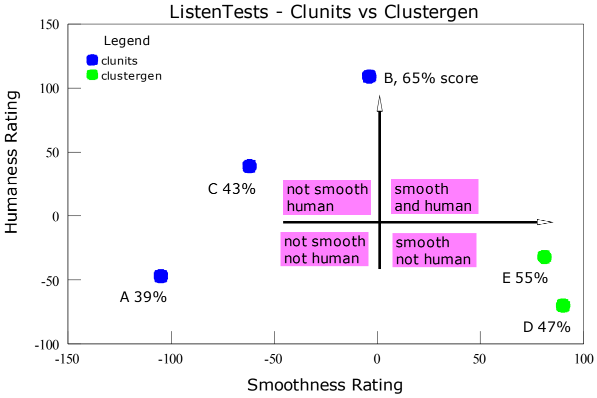

Figure 1. Ratings of the five candidate systems on smoothness versus humanness scales. The CLUNITS points are on the left hand side (not smooth). The CLUSTERGEN points are in the lower right quadrant (smooth but not human-like).

As expected, the CLUSTERGEN voices are judged much smoother than the CLUNITS voices. Incorporating voicing/ unvoicing degraded smoothness by 9 points but improved humanness by 38 points. The baseline CLUNITS voice was judged neither smooth nor human-like. The CLUNITS voice B with contextual name binding was judged most human-like of the five, and moderately smooth. None of our systems secured a spot in the coveted upper right hand quadrant – smooth and human-like.

5. Build3 – Multilevel Unit Selection

Given the preference for CLUNITS, but also the limitation that some sentences could not be synthesized due to missing units, we were reminded of an existing script in the Festvox tools called build3. This was originally developed in 2000, but never fully explained in any other publication.

In build3, 3 different unit selection clunit synthesizes are built on the same data. At each level, different binding applies. In the top level, we bound each phone to both its next and previous phones (cf triphones), the next synthesizer bound its units only to the previous phone (cf biphones), and the final used just the phone itself. In order to be used, it was required that a unit had at least five examples.

At synthesis time, the first (triphone) synthesizer was attempted. However, if no units were found, the biphone synthesizer was used, and if no biphone units were available, the phone synthesis was used. For most part, triphone units existed, but for cases where they did not, we have a well-defined back-off strategy.

Further, before treating them as bi- or triphones, we bound all vowels with stress, and all consonants with onset/coda. A further binding was made after listening experiments, where each phone followed by a "pau" was marked as such, in order to make stronger phrase final choices.

6. Build3 – Tuning

We also adopted a number of other tuning techniques that are optionally available in the CLUNITS tools.

Duration pruning was used to remove all segments from the training set whose actual duration differed from its predicted duration by more than 1 standard deviation. This removed about 10% of the units, but made the synthesis flow sound better, though it likely also reduced its naturalness.

Additionally, a substantial amount of time was used to individually tune the join weights of the system. We incrementally tested different weights in the standard system to find optimal relative weighting values for F0, Cepstral, and other parameters by listening to synthesized test sets. This is quite expensive from a human aspect, but did substantially improve the quality of the system. We would like to claim that our other structure pruning most improved the system, but actually the low-level weight tuning by hand had more of a contribution.

7. Silence Insertion

Following our experience in speech-to-speech machine translation, we are aware that if the TTS input text is more complex than expected, the TTS system will typically still produce it at speed that is appropriate for less complex material. Specifically in S2SMT the output of the MT may not even be grammatically fluent, but the synthesis still reads it as if it is fluent, making the resulting speech much harder to understand [7]. Therefore, following some of our other work, we investigated adding in additional breaks to the text depending on the calculated complexity of the text itself. This was a deliberate attempt to improve scores on SUS sentences. Though this did make more of a difference for SUS sentences than others, the technique was applied to all test sentences, not just the SUS sentences. Using a hand optimized weight function of unigram, bigram, and mutual information scores for each word, tuned on test sentences from Blizzard 2005 [8], we added in silence when it was felt the understandability would be improved.

In spite of what we think is a useful idea, the overall Blizzard results do not seem to back this up. Our relative position in the pack remains the same in both the MOS scores and the SUS scores.

8. Conclusion

The CMU entry was letter M, which performed about sixth overall. This is a substantial improvement from our 2007 position. The system differences are really fundamentally due to taking much more care in labeling the data, and removing data that we were not confident about.

Probably the most significant payback from the time we spent on building our voices was on careful hand tuning of the weights. We were disappointed that such low level time-intensive, database specific work was the most important thing in improving our synthesis. We would have preferred that new general techniques produced important advances, but our experience was that careful hand tuning is still significant in building good unit selection synthesis voices.

9. References

[1] Alan W Black and Paul Taylor, "Automatically clustering similar units for unit selection in speech synthesis," in *Proceedings of Eurospeech 97*, Rhodes, Greece, vol. 2, pp, 601-604, 1997.

[2] Alan W Black, "CLUSTERGEN: A Statistical Parametric Synthesizer using Trajectory Modeling," *Interspeech 2006 - ICSLP*, Pittsburgh, PA, 2006.

[3] Kishore Prahallad, Arthur R Toth, and Alan W Black, "Automatic Building of Synthetic Voices from Large Multi-Paragraph Speech Databases," *Interspeech 2007*, Antwerp, Belgium, 2007.

[4] Susan Fitt and Stephen Isard, "Synthesis of regional English using a keyword lexicon," in *Proc. Eurospeech '99*, Budapest, Hungary, vol. 2, pp. 823-826, 1999.

[5] Gopala Krishna Anumanchipalli, Kishore Prahallad and Alan W Black, "Significance of Early Tagged Contextual Graphemes in Grapheme Based Speech Synthesis and Recognition Systems," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2008.

[6] Hunter, D. MM Algorithms for Generalized Bradley-Terry Models, *The Annals of Statistics*, vol. 32, no. 1, pp. 384-406, 2004.

[7] Tomokiyo, L., Peterson, K., Black, A., and Lenzo, K. "Intelligibility of Machine Translation Output in Speech Synthesis," *Interspeech 2006 - ICSLP*, Pittsburgh, PA, 2006.

[8] Alan W Black and Keiichi Tokuda, "Blizzard Challenge -- 2005: Evaluating corpus-based speech synthesis on common datasets," *Interspeech 2005*, Lisbon, Portugal, 2005.