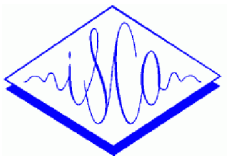
ISCA Archive
http://www.isca-speech.org/archive



INTERSPEECH 2012
ISCA's 13th Annual Conference
Portland, OR, USA
September 9-13, 2012

Proper Name Splicing in Computer Games with TTS

Blaise Potard, Matthew P. Aylett, Christopher J. Pidcock

CereProc Ltd., Edinburgh, UK

blaise@cereproc.com

Abstract

Building high quality synthesis systems with open domain vocabulary and a small audio database is a challenging problem, even when the targeted application is well constrained. Monophone unit concatenation (as opposed to diphone) is an approach that can compensate for the poor unit coverage that a small database implies. However, joining at phone boundaries is a delicate task that requires accurate targeting. In this paper, we present an automatically trained targeting system based on the parametric synthesiser HTS, and compare it to a concatenative monophone system and a baseline concatenative diphone system. We apply a novel evaluation methodology which includes a qualitative component, and allows for fast incremental development of synthesis systems. Preliminary results show that although the hybrid system performed significantly more poorly on out of database items, it is less affected by segmentation errors than the monophone system.

Index Terms: hybrid speech synthesis, unit selection, evaluation of TTS systems

1. Introduction

In speech synthesis applications, the domain of dynamic content can vary greatly. Compare, for example, an application that reads out news articles compared to one which varies the proper name in a series of sentences. In one we have a challenging dynamic environment with different prosodic contents, open ended vocabulary and significant text normalisation problems. In the other, we have almost fixed content with a homogeneous prosody, but also open ended vocabulary for the proper names. In addition we have the problem of avoiding any sharp contrasts between pre-recorded speech and the synthesised proper name slots.

For an open domain speech synthesis voice we are prepared to record a large corpus to create the voice. For a more limited domain application, such as slot filling, we would like to achieve a better efficiency in terms of recording time.

The required quality of speech synthesis, as well as the tolerance for different types of errors, can also vary with the type of application. In this paper we focus on the proper name slot filling problem, with the caveat that it must produce extremely high quality output for as small a recorded corpora as possible. This requirement is driven by the intended application domain of computer games, where production standards are high and *any* speech synthesis artifact is unacceptable. Few researchers have developed systems targeting this particular problem[1], and even fewer have evaluated their performances[2].

We will outline our strategy for producing a 'smart splicing' synthesiser for proper names, evaluate two different systems created for this application, and finally discuss the implications of our findings with regards to a wider application area.

1.1. Concatenation vs Parametric Generation

The first question to address is the choice of a concatenative or parametric synthesis paradigm. The CereVoice engine supports both a concatenative system [3] as well as an implementation of HTS [4]. HTS has shown to be robust and effective for synthesis based on small corpora[5], whereas unit selection systems typically require upwards of 300k phonemes to avoid data sparsity. This is due to the requirement to cover both phonemic and prosodic contexts, which leads to an explosion of required contexts. The main problem with using HTS in this application was that the vocoded sound of the voice was not acceptable to our customers. In computer games a sense of presence, naturalness and character are paramount. Much current research work is focused on improving the vocoded sound of parametric synthesis (e.g. [6]), but for our purposes and with the current state of parametric synthesis systems, a concatenative approach was the only option. Concatenation also offered the advantage of being able to lift carrier phrases from the database with very little modification to our existing system.

1.2. Monophones vs Diphones

The second question surrounds the base unit of the system. Within unit selection, diphone systems have dominated English TTS, whereas monophone systems have often been used for Asian tone languages such as Mandarin. CereVoice supports both a monophone based and a diphone based concatenative approach.

On the one hand, diphones have the advantage of offering a more homogeneous location for joining (in English). Additionally, diphones are often more robust with regards to segmentation errors. However, they require a larger audio database to ensure data coverage.

On the other hand, recent work by USTC[7] has shown that a monophone based approach in unit selection can be effective, especially for smaller databases. In that paper, a *hybrid* approach was adopted where a parametric approach was used to model speech and used as a basis for the unit selection algorithm. Other hybrid approaches include Pollet and Breen[8], Taylor[9], and Tiomkin[10].

For the proper name slot filling application, a monophone approach appears attractive for the following reasons:

1. The carrier phrase will be lifted directly from the database, and many proper names will be covered directly by the recording. Hence the amount of concatenation required should be small, meaning artifacts caused by segmentation errors will be minimised.
2. Even given a restricted prosodic context, phoneme variation in proper names is open ended. Therefore it is not possible to cover all required phonemic contexts using diphones in a short recording session.

3. The CereVoice HTS system uses context based phones, and thus a monophone based concatenation system would allow a hybrid approach, with HTS models being used to select candidate units.

1.3. Unit Selection vs Hybrid Targeting

One significant problem for unit selection systems is the requirement of tuning weights in the target and join cost functions[11]. The hybrid approach offers a more systematic solution to this problem. HTS selects features for modelling based on a tree of feature questions created to reduce minimum description length (MDL). Thus the features which contribute the most to acoustic variation are automatically given more importance in this tree structure. The resulting model gives a relatively small set of model states that can be used to select units.

With the name-splicing application in mind, CereProc developed a hybrid TTS system, using HTS generated phones as the targets for the unit selection. In the CereVoice system this reduced the set of features used in the target cost function from 26 to 16, making weight tuning a simpler process.

In addition, the states chosen by the HTS system implicitly model a spectral target. Such a target is not common in standard unit selection systems. This means that spectral context effects could both be modelled and selected based on HTS state sequences. For a monophone based concatenative system, such context could implicitly model systematic segmentation errors. Potentially, this may help deal with problems caused by segmentation errors, because provided the issue is systematically modelled it may not cause an error. In other words, a good concatenation may not require a good segmentation, but rather a consistent segmentation.

1.4. Research Questions

Our research questions are as follows:

RQ1: What errors are generated in a monophone system given a state-of-the-art automatic segmentation tuned for a diphone system?

RQ2: Can HTS models be used as a spectral target to reduce these errors?

RQ3: Can HTS models be used to select good candidates in a monophone based unit selection system?

2. Methodology

Data was collected in a 2 hour recording session carried out in a sound proof studio using an amateur voice talent selected for his pleasant voice and ability to read material without error. Data was divided between prompts of mono phrasal proper names e.g. “Jamar, Reid, Eugenio, Lewis, Roosevelt” taken from a list of the top 10k most common US proper names and selected to maximise phonemic coverage, and carrier phrases orientated to a sports game domain e.g. “Here Boys you seen the state of Humberto.”

The recording produced 35 minutes of usable speech of which 21 minutes was phonemic data excluding silences and pauses (12k phones, 4.4k words). Note that this is an extremely small database for an open-ended vocabulary system; as a comparison, in open domain concatenative TTS systems a database comprising 5 hours of audio is considered “very small”[7].

The CereVoice system[3] was configured to concatenate on a monophone basis, with cross correlation used to decide concatenation points in unvoiced transitions. A subset of standard HTS features were used with cost functions and weights extrapolated from the CereVoice diphone architecture, to create the “monophone unit selection voice”.

A single speaker HTS voice was then generated using straight analysis, 39 Cepstral order (bark by Julius scale), a 5ms frame rate and 7 state models. The sampling rate was 16KHz with an alpha of 0.58 (see [4] for HTS setup and training details).

The hybrid unit selection voice was created with features extracted for each of the five MCEP, f0 and aperiodic energy state sequences used to label the input data during HTS training. In this preliminary study, a very simple target cost architecture was used, with each state contributing the same cost if a candidate unit is non-matching and zero if matching. For the beginning and end state the cost was multiplied by 10 to encourage better spectral matching at join points.

The segmentation from the monophone unit selection voice was used, rather than the HTS training system alignment. This ensured that both systems were using the same units, for easier comparisons.

Both voices used identical join costs based on spectral and f0 values at the phone boundaries. The only difference between the hybrid and monophone system was the target cost function.

Finally, a third voice was created using our standard diphone system.

2.1. Experimental Stimuli

Twenty sentences were generated using ten carrier phrases present in the database. In ten of these sentences the final proper name was replaced with another in-database version from a different carrier phrase context. In the other ten sentences the final name was replaced with a proper name not present in the database.

Each sentence was synthesised using the monophone unit selection voice, the hybrid system, and the diphone system.

2.2. Evaluation

Our evaluation formed part of a live commercial project. In this context, evaluation was geared more for diagnosis than for assessing final quality. In general, a 5 point MOS scale judging naturalness can give a good impression of the overall quality of a system, but is poor at indicating where and what problems have occurred. We instead carried out a two part in depth listening test:

Word-by-word scoring: Each word in the sentence is given a score of 0 (good) 1 (synthesis artifact) 2 (critical synthesis artifact). The sentence can be listened to as many times as required.

AB comparison test: A listener hears a pair of sentences from two systems in random order, and a preference from -2 (prefer A) to 2 (prefer B) is given. The sentence can be listened to as many times as required.

For this preliminary evaluation we used 5 subjects, all highly qualified speech synthesis engineers, with over a decade of experience in the field. This *expert* evaluation approach allows a fast turn around of materials, and fast diagnosis of system problems.

We then applied a qualitative evaluation to the 8 sentences which showed the biggest error difference between the monophone and hybrid systems. This was carried out using in house diagnostic tools to investigate problem units and the segmentation used to produce them.

3. Results

Figure 1 shows the average word error scores for the carrier phrase and the proper name. The proper name result is split by whether the name was in the database or not. As we might expect, there were more errors for names outside the database. This effect is especially marked for the hybrid system. The diphone system is significantly worse than the other two systems for names within the database.

Figure 2 shows results for AB comparison tests for sentences where the proper name was in the database and sentences where the proper name was not in the database. Subjects showed a significant preference for the monophone system when the proper name was not in the database (sign test, $n = 100, p < 0.01$). Subjects also showed a significant preference for the monophone and hybrid systems over the diphone one for results where the name was in the database (sign test, $n = 100, p < 0.001$). Note that these results are consistent with the scores observed in the word-by-word test.

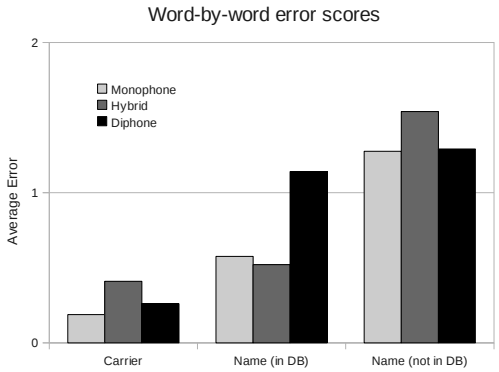


Figure 1: *Average results of word-by-word scoring. Each word was scored 0 (good) 1 (synthesis artifact) 2 (critical synthesis artifact). For the carrier phrase the highest score (excepting the name) was used to represent the phrase. Results are shown for words in the carrier phrase only, proper names within database, and proper names not in the database.*

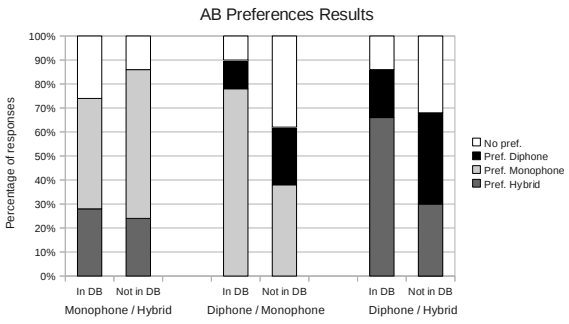


Figure 2: *Results for AB comparison test. For proper names in the database, there is a significant preference for the monophone and hybrid systems over the diphone one. For proper names outside the database, there is a significant preference for the monophone system over the hybrid one.*

3.1. Qualitative Evaluation

Eight sentences were investigated in depth to compare monophone and hybrid systems more precisely. Four sentences showed a strong preference for the hybrid system, the other four for the monophone system.

Table 1 shows the results of the detailed analysis. A number of the errors are caused by segmentation problems, and these errors appear more frequently on the monophone system. Figure 3 gives an example of how a small segmentation error in a phone based system can cause an artifact during concatenation: the small part of the vowel appended to the end of the fricative will cause a problem during synthesis. In contrast, in a diphone system this error would not result in an artifact.

In the hybrid system, the problems are caused more often by the poor selection of a unit. For example, one unit was selected with an inappropriate context. In the next section we will discuss why these problems occurred, and how we would expect to solve them.

Table 1: Qualitative Analysis. The first column "S." shows the system where the error occurred ("M": "Monophone", "H": "Hybrid").

S.	Word	Description	Category
M.	Scott	Elision of /t/ in 'past' caused segmentation error of initial /s/ in 'Scott'	segmentation
M.	Helen	Missing /h/ from 'Humberto' together with bad segmentation between /eh/ and /l/ in 'Roosevelt' caused loss of initial syllable.	segmentation
M.	Rubbish	Final /sh/ segmented with 25ms of following /ow/ in 'show' causing artifact before following vowel.	segmentation
M.	Stephen	Initial /s/ in Stephen segmented with 44ms of proceeding /eh/ causing artifact word initial.	segmentation
H.	Jack	Inaudible /k/. The /k/ was almost totally elided in its original context.	selection
H.	tonight	Final /t/ is an /l/ and 't' was not pronounced by the speaker in the word "Roosevelt".	segmentation
H.	Claire	Bad /l/ is preceded by a stop, so sounds like /gl/.	selection
H.	Gary	Bad /eh/ to /r/ join. /r/ does not have the right pre-context, it is preceded by an unvoiced fricative in the original word.	selection

4. Discussion

None of the systems performed sufficiently well to meet the project goals in their current configurations.

Unsurprisingly, the diphone system performed the worst, in particular when the proper name was in the database. Surprisingly, when the proper name was *not* in the database, it did not perform worse than the other systems.

The results for the monophone system were the most promising in that approximately 50% of the sentences were

judged without error. The fact that errors were noted within the carrier phrases, which were entirely present in the database, suggests that changes are required to the weights compared to a standard open domain system. A higher join cost is likely to lift longer sections from the database and reduce these errors.

Of more concern is the results from the qualitative assessment of the poor monophone sentences. Figure 3 gives a key insight into the problem. This segmentation error would not cause a problem in a diphone system. In the monophone system, with no spectral measure available to select units, it is very likely to cause an artifact *even if it can be joined smoothly to a following unit*. The problem is that the join cost does not indicate whether a sound is appropriate, merely whether one sound can be smoothly concatenated to another. The result in the monophone system was a series of unwanted artifacts that sounded less like concatenation errors than errors in the speech content.

One option would be to improve the segmentation at these boundaries. However, any automatic algorithm will still make errors, particularly with a small audio database.

It is here that the hybrid system shows the most promise. The hybrid system should be able to avoid these problem units, because the initial and final states produced during HTS alignment should match the HTS target states requested during synthesis.

Unfortunately, results from the AB test showed a significant degradation in quality for the hybrid system, especially when the proper name was not in the database.

This degradation can be ascribed to two issues:

1. The hybrid target cost function allowed for no back off. If no unit matching the state sequence could be found, then all other units were graded equally poorly. In the USTC system the probability of the model aligning with the chosen unit can be used as a parametric cost (rather than the categorical costs we used in this system). Such a parametric cost based on a distance metric between states would be likely to pick an 'adequate' alternative unit when a 'good' candidate was not available.
2. As stated in the methodology, the segmentation used in the hybrid system was not the same as the segmentation generated by the HTS voice. For many phones this is not a critical issue, and preliminary tests suggested the two segmentations were similar. However this is less the case in segmentation between liquids such as /l,r/ and vowels. For a diphone system these boundaries can be almost arbitrary. Inspection of the error in 'Roosevelt' showed that the HTS segmentation was more appropriate and balanced between the /eh/ and /l/. However, the elided /t/ caused equal problems with both systems.

Results from the evaluation give support to the approach of using word-by-word scoring together with AB testing and a post qualitative assessment: the results of the evaluation give us a clear idea of the elements of the system that need to be modified to improve the speech output.

5. Conclusion

To address our research questions we can conclude the following: Segmentation errors at voiced/unvoiced phone boundaries are the biggest problem when moving from a diphone to monophone concatenative system. HTS modelling can reduce this problem by using the model numbers as categorical targets. However, for the models to select good candidates they require a more finely based target cost function.

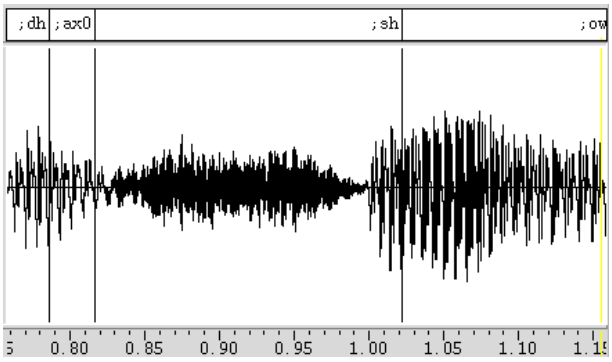


Figure 3: *Segmentation error between /sh/ and /ow/ in the word 'show'.*

More and more frequently, we find that customers want speech synthesis to solve different problems than the classical task of TTS – given some unknown text, speak it correctly. Speech synthesis systems also have to be speech editing systems. In the mobile age, and in particular in video games for mobile platforms, the requirement for smart speech splicing, and for using speech synthesis in constrained – but not closed – domains, is growing. This work presents an approach that can be adapted to meet these challenging requirements.

6. References

[1] R. D. et al, "Phrase splicing and variable substitution using the ibm trainable speech synthesis system," in *ICASSP*, 1999, pp. 373–376.

[2] W. Hamza and J. Petrelli, "Combining the flexibility of speech synthesis with the naturalness of pre-recorded audio: A comparison of two approaches to phrase-splicing tts," in *INTERSPEECH*, 2005.

[3] M. P. Aylett and C. J. Pidcock, "The cerevoice characterful speech synthesiser sdk," in *AISB*, 2007, pp. 174–8.

[4] J. Yamagishi, H. Zen, T. Toda, and K. Tokuda, "Speaker-independent HMM-based speech synthesis system – HTS-2007 system for the Blizzard Challenge 2007," in *Proc. Blizzard Challenge 2007*, Aug. 2007.

[5] H. Zen and T. Toda, "An overview of nitech hmm-based speech synthesis system for blizzard challenge 2005," in *Blizzard Challenge*, 2005.

[6] T. Drugman, G. Wilfart, and T. Dutoit, "A deterministic plus stochastic model of the residual signal for improved parametric speech synthesis," in *INTERSPEECH*, 2009, pp. 1779–1782.

[7] Y. Jiang, Z.-H. Ling, M. Lei, C.-C. Wang, L. Heng, Y. Hu, L.-R. Dai, and R.-H. Wang, "The ustc system for blizzard challenge 2010," in *Blizzard Challenge*, 2010.

[8] V. Pollet and A. Breen, "Synthesis by generation and concatenation of multiform segments," in *INTERSPEECH*, 2008, pp. 1825–28.

[9] P. taylor, "Unifying unit selection and hidden markov model speech synthesis," in *INTERSPEECH*, 2006.

[10] S. Tiomkin, "A segment-wise hybrid approach for improved quality test-to-speech synthesis," Ph.D. dissertation, Israel Institute of technology, 2009.

[11] A. Hunt and A. Black, "Unit selection in a concatenative speech synthesis system using a large speech database," in *ICASSP*, 1996.