INTERSPEECH 2009 BRIGHTON

Speech synthesis without a phone inventory

Matthew P. Aylett[1,2], *Simon King*[1], *Junichi Yamagishi*[1]

[1]Centre for Speech Technology Research, University of Edinburgh, UK
[2]Cereproc Ltd. Edinburgh, UK
matthewa@inf.ed.ac.uk

Abstract

In speech synthesis the unit inventory is decided using phonological and phonetic expertise. This process is resource intensive and potentially sub-optimal. In this paper we investigate how acoustic clustering, together with lexicon constraints, can be used to build a self-organised inventory. Six English speech synthesis systems were built using two frameworks, unit selection and parametric HTS for three inventory conditions: 1) a traditional phone set, 2) a system using orthographic units, and 3) a self-organised inventory. A listening test showed a strong preference for the classic system, and for the orthographic system over the self-organised system. Results also varied by letter to sound complexity and database coverage. This suggests the self-organised approach failed to generalise pronunciation as well as introducing noise above and beyond that caused by orthographic sound mismatch.

Index Terms: speech synthesis, unit selection, parametric synthesis, phone inventory, orthographic synthesis

10.21437/Interspeech.2009-598

1. Introduction

In unit selection and parametric speech synthesis, the inventory of units is decided by inspection and on the basis of phonological and phonetic expertise. The ephone (or emergent phone) project is investigating how self organisation techniques can be applied to build an inventory based on the constraints of a synthesis lexicon together with acoustic observations.

An emergent phone system would allow the rapid development of synthesis systems for languages with a limited body of phonetic and phonological analysis, and for non-standard accents. In addition, such a system could potentially help solve a set of problems commonly associated with unit selection speech synthesis which are currently approached with the use of heuristics and ad-hoc rules (See [1]). By removing ad-hoc rules, and replacing them with machine learning techniques, we will be able to make the unit selection approach to synthesis more formalised and facilitate unit selection/parametric hybrid approaches[1].

In this paper we report on results of using an emergent phone inventory in both a unit selection [2] and a parametric [3] synthesis system. These systems were compared to classic systems, where a conventional phone inventory was used, and systems where orthography alone was used as the unit inventory.

The creation of an emergent phone set can be seen in part as an attempt to merge letters to sound (LTS) techniques with self organisation based on the speech acoustics. In part, an LTS system reflects lexical constraints. For example the same word is generally (although not always) pronounced in the same way, the same series of letters in different words has a higher than random chance of having the same pronunciation. In contrast, acoustic constraints vary between the absolute (does this sound the same?) to relative (given the context, is this section of acoustics the same 'thing' as another section of acoustics). Unit selection and parametric synthesis algorithms have addressed the issue of acoustic constraints by using a classic phone set and by adding rules which enforce or bias a context match. This suggests two alternative strategies for disposing of the classic phone set within synthesis:

1. Use the orthography directly as units and depend more heavily upon the synthesis system's use of context to resolve irregular pronunciation.
2. Use the lexicon and acoustics of a database of input speech to cluster units into an emergent set. Then use standard LTS approaches to generate the pronunciation of unseen words.

Central to any evaluation of these systems is LTS irregularity. This can be divided into two interrelated factors, firstly, the extent the same orthographic sequence matches an identical sound sequence, and secondly, the extent the pronunciation of an unseen word can be deduced knowing these mappings. Therefore, in our evaluation we control for LTS complexity and database coverage.

2. Overview of the systems

2.1. Classic systems

The speech data consists of 2098 utterances, (189k phones, 5.4 hours) of citation speech recorded in a studio environment from a single female English speaker with a received pronunciation (RP) accent. The lexicon was transcribed using the RP MRPA phone set which contains a total of 44 phones (20 vowels).

Features extracted consisted of symbolic features (phone context, word context, syllable context, phrase context etc.) and parametric features (F0 values, LSP spectral parameters and energy). In the unit selection system [2], these features were used to calculate target and join costs for a viterbi search. The standard algorithm was modified to prevent contiguous database joins (normally a preferred outcome), to force the system to test the generalisation of the pronunciation system.

The parametric voice was constructed using Cereproc's implementation of the HTS 2007 speaker dependent system [3]. Only symbolic features were used but contained a wider variety of features than that used in unit selection including syllabic position in phrase, word position in phrase etc. Questions used to cluster models by phonetype were automatically generated by dividing the phones into sets based on distinctive features which included: syllabic, voiced, plosive, fricative, nasal, liquid. The speech was spectrally analysed and synthesised using STRAIGHT [4].

2.2. Orthographic systems

In languages where there is a close relationship between orthography and pronunciation it is possible to build synthesis systems without a phone set but to use the letters instead. This is not a simple task in English where the mapping is complex and irregular. A pre-requisite for both a parametric and unit selection system is the ability to segment the data accurately into the sound units. In [1] we found that an orthographic segmentation of a single speakers data produced comparative word boundary results to that of conventional segmentation. However in order to accomplish this, *extra* units were added beyond the 26 letters of the English alphabet. These extra units allowed word boundary information, common words, and common letter sequences to be included as separate units. The final set consisted of 226 units. The segmentation at word boundaries matched a classic segmentation relatively closely (RMSE 11ms, 13% insertion rate, 19% deletion rate).

Only two features were used to describe these units, voiced and syllabic. These were calculated by noting voicing across each example of a unit. If more than 66% of the frames for a unit type were voiced the feature was set to true. The syllabic feature was calculated using the SLPA toolkit [5], where energy, f0 and a peak picking algorithm were used to place syllable nuclei positions. If a unit type contained a majority of these nuclei positions it was marked as syllabic.

A lexicon was then constructed using these units and used in the conventional voice creation system for both unit selection and parametric synthesis. In effect the output of the process is a lexicon, which contains all words present in the acoustic database. Thus, final segmentation and (for the unit selection system) individual word pronunciations, are free to alter during voice building.

2.3. Emergent phone system

The data was first segmented into unmarked emergent phones using dynamic time warping (DTW) to find common repeating sections of audio. This algorithm was based on [6] and segmented the speech into a set of units based purely on acoustic information (see [1] for more detail of this approach). An orthographic segmentation obtained from the orthographic system was then used to force word boundaries and silence/pause locations onto the self organised segmentation. This produced a final segmentation of 246k units (1.3 times the number of classic phone units).

The emergent system used the orthographic segmentation to bootstrap categories. Orthographic units were mapped onto the nearest emergent phones, and units which accounted for over 1% of the data were then used as a basis for emergent phone categories. The data was remapped into 3 dimensional space using isomap [7] and the DTW algorithm used to segment as a distance metric, then xmeans (an efficient implementation of kmeans using BIC to decide number of clusters [8]), was used to further divide the initial categories. All units remaining were allocated to the nearest cluster based on the DTW metric. The result was a set of 256 emergent units. As with orthographic units, voicing and syllabic features were determined, and a lexicon was created using these units.

A lexicon for unseen words is then generated by using the database lexicon to build a LTS system. In this experiment a system based on learning joint multigrams was used. New pronunciations for unseen words were then generated using this model[1]. The system was also used to replace initial pronunciations to make the lexicon more homogeneous but not used to score LTS complexity (see section 3.1).

3. Evaluation

3.1. Experimental design

18 sentences were selected from 10000 sentences with 6 to 10 words taken from web news output.The sentences were selected on the basis of two factors:

How many of the words were present in the initial speech database All sentences with out of vocabulary words (not in Cereproc's standard 128k word RP lexicon) were removed. Two sets of sentences were then generated, one where all words in each sentence were in the initial audio database and one where as many out of database words as possible were present.

The extent the words pronunciation and orthography had a simple LTS relationship (LTS complexity) LTS complexity was calculated on the basis of the inverse of the score for carrying out a DTW alignment as proposed in [9]. Briefly, expectation maximisation is used to determine a frequency table mapping letters to phones, this is used to score a DTW alignment together with an insertion/deletion penalty. For example, the most LTS complex word was the pronunciation of the word 'w', other hard examples include 'shoe', 'whey', easier examples include words such as 'ants', 'intend' where each letter maps directly onto it's most likely phone.

In addition, word lists were extracted based on the same factors and grouped by nouns, adjectives and verbs in order to semi-automatically create semantically unpredictable sentences (SUS), for intelligibility evaluation.

Thus, for every audio file generated the following factors independent factors were assigned:

System/Phone set: HTS, Unit Selection (USL) and the inventory type: Classic (PHN), Orthographic(ORT), Ephone (EPH)

Database: Within audio database (INDB), not in audio database (EXDB)

LTS complexity: Easy (ESY), Medium (MED), Hard (HRD)

21 subjects listened to the audio files using a web interface. They graded each sentence using a scale from 1 (Not natural) to 5 (Completely natural)[2]. A latin square design was used to ensure no subject heard the same sentence twice.

In addition each subject was asked to type a transcription of each SUS sentence which had the following form *'The heated breadth exudes the jazzed caraway'.* into an input box.

3.2. Results

3.2.1. Analysis

Results for each utterance were converted into a mean opinion score (MOS). Results using MOS scores should be treated with care as there is a strong argument that the underlying subject data should not be treated as parametric data. However MOS is a default standard in speech synthesis and using MOS allows a multifactor analysis of the data using a grouped ANOVA analysis. Although MOS data is rarely Gaussian, an ANOVA analysis

[1]Thanks to Dong Wang at CSTR for the use of his prototype system.

[2]see `http://www.cogsci.ed.ac.uk/~matthewa/interspeech2009wavs` for all materials used in the experiment encoded as mp3s

is acceptable based on the sampling theorem providing each cell has sufficient data points (commonly 10 or above).

All three independent factors were significant (n = 126): System/Phoneset (F=47.9, $p < 0.001$, df 5), Database (F=7.3, $p < 0.01$, df 1), LTS complexity (F=7.3, $p < 0.005$, df 2). In addition two interactions were significant: System/Phoneset * LTS complexity (F=2.7, $p < 0.005$, df 10) and Database * LTS complexity (F=3.6, $p < 0.005$, df 2).

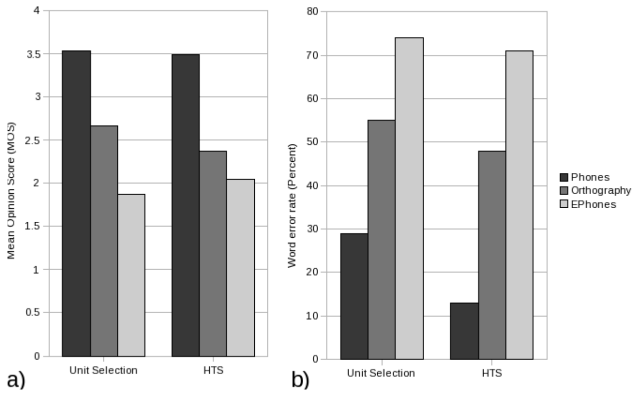

Figure 1: *a) Mean opinion scores (MOS) of systems and phone inventories. There are no significant differences between HTS and Unit Selection systems. All phone inventories are significant ($p < 0.001$), except between HTS Orthography and HTS ephones. b) Word error rates in semantically unpredictable sentences (SUS) by system and phone inventory.*

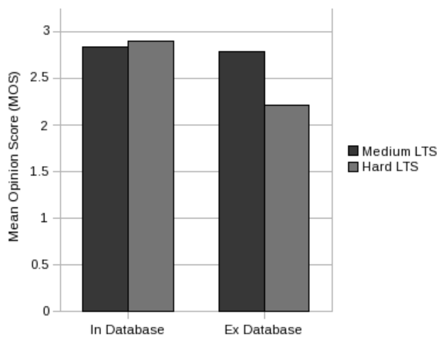

Figure 2: *The interaction between database coverage and LTS complexity.*

Figure 1a shows the MOS results for each system/phoneset. There is a clear preference for the classic systems. Orthographic systems perform better than ephone systems (although the difference for parametric synthesis is non-significant.). Results for the database factor were as expected, sentences which contained many words not in the original database performed worse than sentences which contained words from the original database (mean MOS: 2.8 INDB, 2.53 EXDB, $p < 0.005$), However results for the LTS complexity did not follow the expected pattern. The so called easy LTS sentences gave similar results to the hard complexity LTS. In contrast the medium complexity LTS sentences were, as expected, easier than the hard LTS complexity sentences. This could have been caused by 'easy' sentences containing shorter more frequent words which led to hypo-articulation. More work is required to use this LTS measure effectively, however we will present results for the 'hard' and

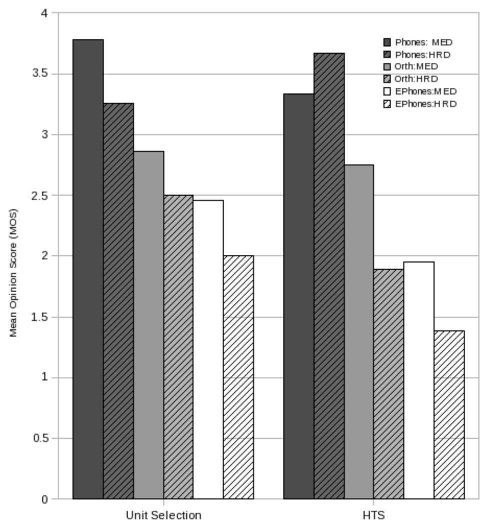

Figure 3: *The interaction between system, phone inventory and LTS complexity.*

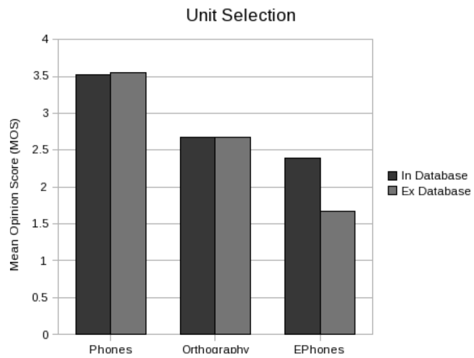

Figure 4: *Effect of database coverage on system for Unit Selection.*

'medium' categories in terms of their interaction with Database and System/Phoneset.

Figure 2 shows the interaction between database and LTS complexity. The effect of LTS complexity appears to come into play most strongly for ex-Database sentences ($p < 0.05$).

Figure 3 shows the interactions between LTS easy and hard categories by system and phone inventory type. A post hoc Tukey test was carried out on this data. Results suggest the power of our analysis is weak over these smaller cell sizes with neighbouring results not showing significant differences (Full results for all 144 post hoc results are not shown due to space limitation). In general a difference of at least 1 MOS indicates a significant difference.

There was no interaction between database and system/phoneset over all data. However a pattern does appear to emerge for the unit selection system. An ANOVA carried out with the LTS factor removed does produce such an interaction with the sole significant result being for the Unit Selection system. This result should be treated with care but may indicate an underlying reason for the poor performance of the ephone system. Figure 4 shows the results by database/system/phoneset for unit selection only. It appears the ephone system has failed to generalise outside the initial acoustic database.

A detailed analysis of the SUS results is beyond the scope of this paper. However it is important to note that results for intelligibility followed a similar pattern to MOS results with perhaps an even more marked difference in system performance. The parametric system achieved generally higher intelligibility scores (a finding in line with previous work. Results for the ephones system were catastrophic see figure 1b.

4. Discussion

Results from this study suggest we have a long way to go in order to replace a classic phone inventory. However it is important to bear in mind that the techniques we have developed in this work also applies to customising current inventories for specific speakers and accents. Looking closely at why both our orthographic and emergent inventories performed poorly gives considerable insight into some key problems within speech synthesis.

1. In some respects it is misleading to look at this work in terms of unit inventories. In reality we are determining a pronunciation lexicon that gives rise to an inventory. Given a database of speech this lexicon can be split clearly into two parts, words present in the database and words which are not present.
2. An initial requirement of any such lexicon is that it can be used to segment speech into words and component sounds that can either be selected by unit selection or modelled by parametric techniques.
3. The traditional view of the *phoneme* is a key issue in such a lexicon. Can a unit be replaced by another unit of the same category without altering the meaning of the word. In effect, do minimal pairs remain valid given the new lexicon.
4. Furthermore, does the lexicon contain context or inventory identity information that can be used to effectively choose (or categorise) units effectively to produce natural speech variation and to convey prosodic variation.

If we look at the lexicon generated by the emergent and orthographic phone system we can make some immediate observations. Both required a very large inventory, in both cases over five times larger than the classic system. For the orthographic system this was required to give acceptable word segmentation. However we did not investigate how many of these units were *required* to give acceptable word segmentation. Such a large inventory will cause problems with unit selection and parametric modelling due to data sparsity *even when the units do represent distinct acoustic items.* Similarly, for the emergent system, the number of unit types proposed by k-means and BIC is too large. Although, as with the orthographic system, this large inventory does not prevent effective word segmentation. The set is large because acoustics differences that are irrelevant to the human ear are modelled as well as those that are not. If the inventory is reduced, the merging of categories *does not represent the human perception of phonological difference*. Thus we have the choice between a large set which does not generalise or a smaller set which conflates categories which are perceived as different by human subjects. Within the emergent system this over fitting of the speech is further exhibited in the average number of pronunciations for the same word within the database. The classic lexicon has approximately 1.3 multiple pronunciations for each word. The initial output from the clustering produces a database lexicon with an average of over 4 pronunciations for each word. In effect almost all instances of an identical word have a different pronunciation. Using LTS to make the lexicon more homogeneous by using a learnt multigram model to regenerate the database lexicon reduces this down to approximately 1.6. However this is still very high.

The key problem, therefore, is that our acoustic clustering, even when using orthographic data as a prior to guide the process, is not representing the speech appropriately. There are two potential reasons for this: firstly the feature space we are using to represent the units does not represent human perception of difference, secondly our clustering process does not cope well given the feature space. Although clustering the acoustics of speech has been part of many systems for many years, in nearly all cases this clustering has been carried out with heavy supervision and more importantly, the clusters are used to model the speech but *not to categorise it*. For example, the use of a multiple mixture Gaussian model in an HMM model.

There has been significant work looking at feature extraction which is more closely aligned with human perception (e.g. [10, 11]. Such a feature space is a critical requirement for a successful ephone system and will form part of future work.

5. Acknowledgements

Support for this research was provided by EPSRC (award number EP/D058139/1).

6. References

[1] M. P. Aylett and S. King, "Single speaker segmentation and inventory selection using dynamic time warping self organization and joint multigram mapping," in *SSW06 ISCA Workshop*, 2007.

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

[3] 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.

[4] H. Kawahara, I. Masuda-Katsuse, and A. Cheveigne, "Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based f0 extraction: possible role of a repetitive structure in sounds," *Speech Communication*, vol. 27, pp. 187–207, 1999.

[5] M. P. Aylett, "Detecting high level structure without lexical information," in *ICASSP*, 2006.

[6] A. Park and J. Glass, "Towards unsupervised pattern discovery in speech," in *IEEE ASRU*, 2005, pp. 53–58.

[7] J. B. Tenenbaum, V. de Silva, and J. C. Langford, "A global geometric framework for nonlinear dimensionality reduction," *Science*, vol. 290 (5500), pp. 2319–23, 2000.

[8] D. Pelleg and D. Baras, "*K* -means with large and noisy constraint sets," in *ECML*, 2007, pp. 674–682.

[9] R. Damper, Y. Marchand, J.-D. Marsters, and A. Bazin, "Aligning letters and phonemes for speech synthesis," in *ISCA SSW5*, 2004, pp. 209–214.

[10] H. Hermansky, D. Ellis, and S. Sharma, "Tandem connectionist feature extraction for conventional hmm systems," in *ICASSP*, 2000, pp. 1635–8.

[11] J. Frankel, M. Magimai-Doss, S. King, K. Livescu, and O. Çetin, "Articulatory feature classifiers trained on 2000 hours of telephone speech," in *Proc. Interspeech*, 2007, pp. 2485–2488.