The CSTR/Cereproc Blizzard Entry 2008: The Inconvenient Data

J. Sebastian Andersson[1], *Leonardo Badino*[1], *Oliver S. Watts*[1], *Matthew P. Aylett*[12]

[1] CSTR, University of Edinburgh, UK
[2]CereProc Ltd, Edinburgh, UK
matthewa@cereproc.com

Abstract

In a commercial system data used for unit selection systems is collected with a heavy emphasis on homogeneous neutral data that has sufficient coverage for the units that will be used in the system. In this years Blizzard entry CSTR and CereProc®present a joint entry where the emphasis has been to explore techniques to deal with data which is not homogeneous (the English entry) and did not have appropriate coverage for a diphone based system (the Mandarin entry where tone/phone combinations were treated as distinct phone categories). In addition, two further problems were addressed, 1) Making use of non-homogeneous data for creating a voice that can realise both expressive and neutral speaking styles (the English entry) 2) Building a unit selection system with no native understanding of the language but depending instead on external native evaluation (the Mandarin Entry).

Index Terms: speech synthesis, unit selection.

1. Introduction

CereVoice®is a unit selection speech synthesis SDK produced by CereProc Ltd., a company founded in late 2005 with a focus on creating characterful synthesis and massively increasing the efficiency of unit selection voice creation. Cereproc enjoys a close relationship with the Center of Speech Technology Research (CSTR) at Edinburgh University, and the CereVoice system is also made available for research use.

The blizzard entry comprised of a Mandarin entry and a full database English RP entry.

1.1. Unit selection without native language skills

Although the Blizzard team had access to a handful of Mandarin speakers for evaluation, none of the team members had any native (or even limited non-native) expertise in Mandarin excepting some knowledge of Mandarin tone structure and phonology. Whereas many Mandarin unit selection systems are syllable based ([1],[2] and [3] among them), CereVoice is a diphone based system. Given the lexical effect of tone on Mandarin a naive approach was taken making each 'phone' dependent on both phone and tone category giving an inventory of 216 phones. Although this approach meant that much contextual information in tone production is retained it also leads to problems with data sparsity. Two approaches were taken to deal with sparsity: 1) phone backoff; 2) post synthesis pitch modification (see sections 3.1, 3.2 and 3.3). In order to make efficient use of the Mandarin listeners we had available, more qualitative data was collected than in a traditional MOS style test (see section 3.4).

1.2. Unit selection with non-homogeneous data

The RP database selected for the Blizzard Challenge contained significant variation in both speaking style and acoustic properties [4] [5]. The variation offers the possibility of a richer more expressive synthetic voice [4] [5], but can also cause problems in terms of inappropriate prosody and concatenation artifacts. The database was however quite large offering the option to leave out more expressive material. Although this may have resulted in more consistent neutral synthetic speech, we decided to harness the prosodic variation of the database and create a voice that can realise various speaking styles and levels of expressiveness. Two main approaches, described in more detail in section 4, was explored for this purpose:

1. **Genre pruning:** where data was marked as having various speech styles or genres, and selection was biased towards using speech from the same genre and prevent inappropriate mixing across genres.
2. **Emphasis:** mark-up of emphasised syllables in the database, where they were only used if requested by the front-end, and similarly to [5], concatenation of emphasised units with non-emphasised units were avoided.

2. Overview of the system

CereVoice is a faster-than-realtime diphone unit selection speech synthesis engine, available for academic and commercial use. The core CereVoice engine is an enhanced synthesis 'back end', written in C for portability to a variety of platforms. The engine does not fit the classical definition of a synthesis back end, as it includes lexicon lookup and letter-to-sound rule modules, see Fig. 1. An XML API defines the input to the engine. The API is based on the principle of a 'spurt' of speech. A spurt is defined as a portion of speech between two pauses.

To simplify the creation of applications based on CereVoice, the core engine is wrapped in higher level languages such as Python using Swig. For example, a simple Python/Tk GUI was written to generate the test sentences for the Blizzard challenge.

The CereVoice engine is agnostic about the 'front end' used to generate spurt XML. CereProc use a modular Python system for text processing. Spurt generation is carried out using a greedy incremental text normaliser. Spurts are subsequently marked up by reduction and homograph taggers to inform the engine of the correct lexical variant dependent on the spurt context.

3. Mandarin Entry

One of the main interests in building a Mandarin voice for the CereVoice TTS system was to investigate the adaptability and flex-

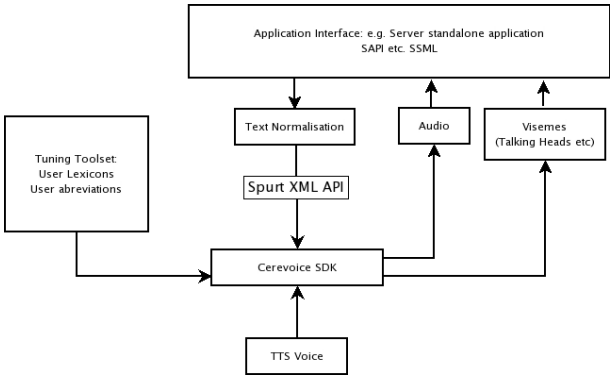

Figure 1: *Overview of the architecture of the CereVoice synthesis system. A key element in the architecture is the separation of text normalisation from the selection part of the system and the use of an XML API.*

	Blizzard data	Blizzard + Extra data
No. of phone types	216	219
No. of diphone types	9004	11806

Table 1: Figures on phone and diphone types.

ibility of the system when dealing with a language far different from English. As a consequence we adopted a "minimum alteration" principle: keep the core elements of our system unchanged as much as possible and apply Mandarin-dedicated modifications only when strictly necessary.

Following this principle we kept the diphone as the basic unit despite most Mandarin speech synthesis systems are syllable based.

Also our definition of Mandarin phone was dictated by our "minimum alteration" principle and the phones were made dependent on both phonetic-articulatory and tone categories.

Our phone definition increased the number of basic phones to 216, and, along with the decision of using a diphone based voice, led to a data sparsity problem. Our first step was to see the extent of that problem: using additional textual data (kindly provided by iFLYTEK) containing the same number of sentences of the Blizzard data, we computed the number of diphone types that were missing in the Blizzard data but were in the additional data. As shown in table 1, 11806 is the overall number of diphone types in the joint data (Blizzard + additional data) and the 23.73 % of them (2802 diphones) was missing from the Blizzard data. We also found that 3 out of 219 phones were missing from the Blizzard data. Preliminary listening tests carried out on an early version of the Mandarin voice, confirmed that missing diphones were the major cause of critical errors in the synthesis. On this basis, we regarded a diphone backoff strategy as the most important step towards an intelligible Mandarin voice.

3.1. Diphone Backoff: phoneme identity

Looking at the Mandarin phone set we noticed that, since it is rich in diphthongs and nasalized vowels, for almost all the vowels the two halves of each vowel have at least one almost identical (in terms of phonetic-articulatory features) counterpart in another

	No. of missing diphone types	No. of missing diphones
No backoff	2802	8229
Backoff: method 1	628	1787
Backoff: method 1+2	211	687

Table 2: Figures on missing diphones

vowel. So, for example, the first half of the diphthong /ai/ shares several phonetic-articulatory features with the two halves of the vowel /a/, while the first half of the nasalized vowel /an/ is very similar to the first half of the vowel /a/ and even more to that of the nasalized vowel /aŋ/ , and so on. As a consequence, we expected that in many cases, swapping a half of a phone (and so a half of a diphone) with a phonetically similar half phone from a different phone would not cause a perceivable degradation in the synthetic speech. We exploited this large availability of interchangeable halfphones (which, unfortunately, does not occur for consonants) by choosing, for each phone, the two closest backoff halfphones of its left half and of its right half. Each vocalic halfphone and its backoff halfphones had matching tone. In general the first backoff halfphone is more similar to the target halfphone than the second backoff halfphone; when the two backoff halfphones were considered equally close to the target halfphones the order of the two backoff halfphones was determined by the frequency of the phone within the Blizzard data: the halfphones belonging to the phone with the highest frequency was chosen as first backoff diphone.

During synthesis, when a diphone was not found in the speech database, its backoff diphone was selected as shown in figure 2. This first backoff strategy reduced the number of missing diphone types found in the additional data from 2802 to 628 (see table 2).

3.2. Diphone Backoff: tone identity

To further reduce the number of missing diphones we introduced a second backoff strategy which was applied when the first strategy failed to find a backoff diphone. In this second method the backoff halfphones differ from the target halfphone by tone only. The priority order of the backoff halfphones is shown in table 3.

In this second method, the backoff diphone is selected looking at three backoff halfphones instead of two but the method of diphone search is very similar to that of figure 2. If the resulting backoff diphone was made up of at least a first (in the priority order) backoff halfphone some post synthesis pitch modification were applied (see section 3.3).

The joint use of the two backoff strategies reduced the number of missing diphone types from 2802 to 211 (see table 2).

3.3. Post pitch modification

The second, part of the strategy used to deal with units missing from the training data was to manipulate the f0 of synthesised waveforms post-synthesis. This allowed the system to back off to units with the correct spectral features but the wrong tone features, the tone then being modified in this post-processing stage. Tests were done to see which Mandarin tones might be converted most successfully into other tones with simple f0 reshaping. The result of these tests was the selection of one "tone backoff"for each of Mandarin tones 1–4. That is, phones with tones 1, 2, 3 and 4 were

Half Phone	
t_l	left target half phone.
$b1_l$	first backoff for left half phone.
$b2_l$	second backoff for left half phone.
t_r	right target half phone.
$b1_r$	first backoff for right half phone.
$b2_r$	second backoff for right half phone.
Search Order	
$t_l\_b1_r$ $bl_l\_t_r$ $t_l\_b2_r$ $b2_l\_t_r$ $bl_l\_b1_r$ $bl_l\_b2_r$ $b2_l\_b1_r$ $b2_l\_b2_r$	

Figure 2: Order of search of backoff diphones. *When a diphone is missing from the speech data the backoff method searches new diphones until it finds a diphone available in the speech database.*
.

Tone of the target halfphone	Tone of the 1st backoff halfphone	Tone of the 2nd backoff halfphone	Tone of the 3rd backoff halfphone
1	2	3	4
2	3	1	4
3	4	2	1
4	3	1	2

Table 3: Priority order of the backoff halfphones for the diphone backoff method based on tone identity. *Tone 1 = high-level tone, tone 2 = rising tone, tone 3 = dipping tone, tone 4 = falling tone*

allowed to back-off to phones with the same spectral characteristics, but with tones 2, 3, 4, 3 respectively. These substitutions were incorporated into the list of allowable back-offs.

This was a simple extension to make to the system as the CereVoice engine allows the inclusion of XML tags in the input spurt specifying the reshaping of the output f0 contour. Pitch modification is carried out using a WSOLA like time stretching algorithm in the time domain, supplied by the freely available LGPL Soundtouch library. When the spurts were prepared, the appropriate tags were added. The procedure followed for incorporating the tags in the spurts is shown in Fig. 3).

To determine the actual f0 shift to specify, the following rules were followed. The f0 could be scaled by 3 values: 1.2, 0.8, or 1 at either end of a syllable. The choice of scaling factor and manipulation point were determined by jointly considering which of the four tone back-offs was being performed, and whether it was the first or second half of the syllable that was backed-off. An added complication was that XML tags must be inserted around whole words, and not around syllables or half-diphones. Thus, the assumption was made that syllables were of uniform length, and "null-entries" specifying that the f0 should be scaled by 1 were included on those parts of words where no back-off was to take place.

This alteration of f0 was based on the naive assumption that a tone value specifies an unaltering f0 contour, not influenced by phonetic, linguistic and discourse contexts. The expectation, however, was not that the target contour would be matched exactly through manipulation, but rather that it would be close enough for a native listener to use top-down processing to perceive the target contour.

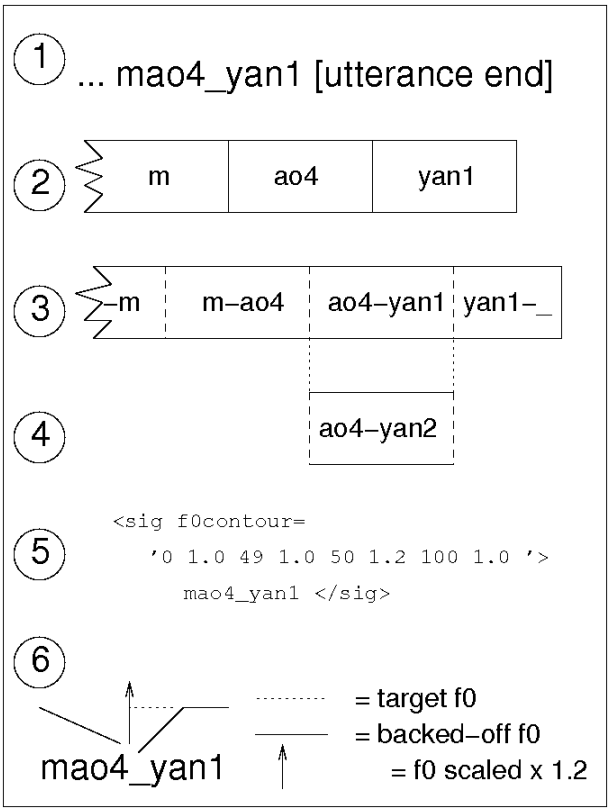

Figure 3: The procedure followed for tone modification.*(1) Input text, (2) Text was parsed into phones (3) Diphones are formed (4) Diphones are checked for requiring backoff, (5) if one of the four "tone back-offs" was detected, then an f0 tag was added, (6) post dsp modification was carried out on the synthesised. uttereance.*

3.4. Qualitative evaluation

In order to make efficient use of the Mandarin listeners we had available we asked them to carry out some short listening tests at different stages of development of the Mandarin voice. A graphical user interface was used to capture word by word errors reported by a native speaker. The listeners signalled the errors they heard by simply mouse-clicking on the piece of text corresponding to the section of speech containing the error. The listeners could choose between two different error scores (1 for a mild error, and 2 for a serious error). By summing the error scores given by all participants for each word we were able to detect the sections of speech containing the most serious errors and to identify their causes. The sum of error scores was also used to obtain a quantitative evaluation of the benefits arising from using the diphone backoff strategies and post pitch modifications described in the previous sections and to discharge other methods that turned out to be unsuccessful. However, this data was not statistically significant due to the small number of test utterances and listeners.

4. English Entry

The English (RP) database selected for the Blizzard Challenge contained a variety of speaking styles and acoustic properties that allowed for more expressive speech synthesis [4] [5], but our ini-

tial voice build revealed that it can also cause considerable problems with inappropriate prosody and concatenation artifacts when different parts of the database are used within the same synthetic utterance.

We decided to utilise the prosodic variation in the Blizzard database to create a synthetic voice that can realise a variety of speaking styles, and levels of expressiveness, and at the same time avoid inappropriate mixing of the different parts of the database. For this purpose we selected the two parts of the database with distinct and different prosodic styles:

- **News:** The *Herald* part of the database representing a news reading style of newspaper sentences. (221min of phonetic material)
- **Fairy-Tale:** The *Carroll* part of the database representing a more expressive fairy-tale reading style of dialogue rich fiction. (84min of phonetic material)

The *Arctic* part of the database (50min of phonetic material) was also included because it represents a prosodic genre roughly between news and fairy-tale: text from fiction, but not read in a "spirited" manner. We left out the *wordlist* part of the Blizzard database because it lacked a natural sentence context and it did not contribute much given the genre pruning described in section 4.2. At synthesis time a simple language model is used to automatically decide the appropriate genre for a particular input sentence.

In addition to the genre pruning we also utilised parts of the Blizzard database to synthesise emphasised words and thus add more expressivity to the voice:

- The prescribed emphasised words in the *Carroll* part [4].
- The emphasised names in carrier sentences [5] (henceforth *Names*).

All the amplitude of the selected data was normalised. In addition limited companding was also applied to further reduce amplitude variation. All the selected data was segmented with forced alignment from a flat start using the HTK tool-kit [6]. A few sentences with faulty transcription or cut-off audio was corrected or removed.

4.1. Speech Synthesis Speaking Styles

Previous Blizzard Challenges [7] [8] have already showed that many systems achieved good results on generating neutral sounding speech. Synthetic voices, whether unit selection or HMM-based, that can realise various speaking styles are mainly focused on synthesising different emotions or voice qualities; e.g. [9] [10] [11]. But we believe that it's equally important to focus on realising different speech genres, such as news, fairy-tales or conversation.

The dataset selected for the Blizzard Challenge have previously been used to realise boundary tones and emphasis in unit selection speech synthesis and showed that it improved perceptual impressions [4] [5]. The *Carroll* data was recorded because it contained a variety of sentence types (e.g. questions, exclamations, quotations) and an existing mark-up of emphasised words and the voice talent was instructed to read *"in a spirited manner"* [4, p.2]. Some news sentences were also recorded to boost coverage [4] and additional news and *Arctic* was recorded for [5].

Whereas [4] [5] focused on local prosody in terms of recording particular diphone units to realise appropriate boundary tones and emphasise words, we focused on global prosody, the clear difference in speaking style between the *Carroll* and news (*Herald*) parts of the database, and create a voice with distinct speaking styles.

4.2. Genre Pruning

The selected parts of the Blizzard database (*Carroll*, *Herald*, *Arctic* and *Names*) were tagged with different genre tags. If a specific genre was requested in the front-end at synthesis time the back-end attempted to prune out candidate units with a genre mis-match before the Viterbi search, and hence a pre-selection gave precedence to units with matching genre and resulted in a more coherent prosodic style of a synthesised utterance. If there was less than 50 candidate units with the correct genre, units from other genres were not pruned out. Initial error analysis revealed that half of the concatenation artifacts were vowel joins across genres, on this basis a cost was added to prevent joining of voiced material across genres.

4.2.1. Genre Pruning Discussion

The genre pruning worked well in the sense that very few units from the "wrong" genres are selected, and that we could realise utterances with two distinct speaking styles (if biased in either direction). However the synthesis quality of the *Carroll* genre was less consistent than for *Herald* and resulted in more concatenation artifacts. If the synthesis was carried out without biasing towards any genre the result seemed more stable than if biasing towards *Carroll*, but we also lost out on the more interesting and vivid aspects of the expressive *Carroll* data. A synthesised example of a genre biased and unbiased dialogue from *The Adventures of Tom Sawyer* [12] is available at: `http://homepages.inf.ed.ac.uk/janderss/blizzard2008_examples/`

An informal listening test with three participants on held-out *Carroll* and *Herald* data shows no clear preference for neither genre bias nor non-bias. This test was however done without the language model genre selection described below (section 4.2.2)

4.2.2. Language Model Genre Selection

To automatically select a desired genre from input text a simple language model was applied that counted which part (*Carroll* or *Herald*) contained most of an input sentence's word unigrams and bigrams, where 1 point was awarded for each unigram and 2 points for each bigram, in case this was equal between the two genres the news genre (*Herald*) was chosen because it contains more data than *Carroll*.

At the moment the language model is overly biased towards news, for example in a test on a dialogue rich part of [12] 41% (22/54) of the sentences was biased towards *Carroll*. But 21 out of the 54 sentences got the same score from our language model and hence were biased towards news.

For the sentences in the Blizzard evaluation our language model selected the *Herald* and *Carroll* genres in 85% and 15% respectively, with slightly higher percentages of *Carroll* for the conversational (18%) and novel (21%) and only in 7% of the news test sentences. Showing that neither test set was particularly close to our intended *fairy-tale* genre.

The language model based genre selection could likely be improved in several ways, the most important being to recognise a genre from larger chunks of texts than just per sentence and that it reflects the produced synthetic quality.

4.3. Emphasis

Since the dataset contained material that allowed for realisation of emphasis and would potentially be evaluated in the Blizzard Challenge we tried to utilise the data for this purpose. But we also tagged emphasised words in the database to avoid using them in contexts where emphasis was not explicitly requested in the front-end.

4.3.1. Tagging Emphasis

The Festival [13] utterance files in the Blizzard release contained mark-up of prescribed emphasis that was used to identify emphasised syllables in the *Carroll* and *Names* parts of the Blizzard database. A few of these prescribed emphasis tags was however wrong, the speaker emphasised another more natural word, and those tags were changed manually (prescribed emphasis in capitals), e.g. in the phrases:

roger_0503: *stupid THINGS!*

roger_1124: *much BETTER!*

A subjective analysis suggested that a lot more than the prescribed words were emphasised in *Carroll*, however, we chose not to embark upon any acoustic classifications.

4.3.2. Synthesising Emphasis

To begin with we set our target cost for emphasis very high so that if we had any matching diphone units they would be selected. This revealed that joining in emphasised vowels rarely sounded natural and a high cost was added to avoid splitting emphasised vowels. Using units from the emphasised *Names* rarely resulted in a natural or emphasised sounding word, this was especially true for sentences biased towards *Carroll* since this genre was so expressive that "adding" emphasis from another genre rather seemed to give the impression of reduced emphasis. (actually it mainly sounds "out of character")

The target cost for emphasis was therefore, heuristically, substantially lowered to a level which seemed to give a good trade-off between trying and succeeding in realising emphasis in our synthetic voice. Although this conservative approach meant that we couldn't realise emphasis for arbitrary words, it also meant that *if* emphasised units are selected the resulting speech often sounds emphasised without lowering the naturalness of the utterance.

4.3.3. Emphasis Results

The Blizzard evaluation contained twenty sentences with words that were supposed to be realised as emphasised. These words were marked up with emphasis XML-tags in the front-end. The language model genre selection was applied for these sentences as well and resulted in eleven sentences biased towards *Herald* and nine towards *Carroll*.

A look at the selected units for emphasised words showed that for the nine sentences biased towards *Carroll*, five were realised with emphasised units from the same word type and genre (e.g. the units selected to emphasise *"not"* were taken from an emphasised *"not"*). But only one of the sentences biased towards *Herald* had a word realised with emphasised units, and no units were taken from emphasised *Names*.

5. Blizzard Evaluation Results

The general approach to evaluation and analysis was similar to last year's Blizzard Challenge with a focus on MOS and intelligibility tests [8].

5.1. English Results

The MOS score for all listeners and systems, as provided by the Blizzard Organising Committee are shown in figure 4. In the figure the CSTR-Cereproc entry is represented by letter *K*, natural speech by letter *A* and a unit selection baseline by letter *B*.

The system performed well overall despite the fact that the expressive elements that we focused our development work on were not heavily evaluated.

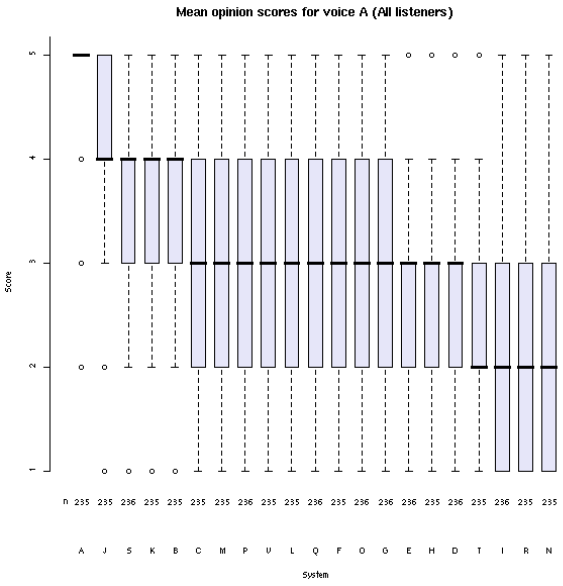

Figure 4: *MOS of the English voices*

5.2. Mandarin Results

The MOS score for all listeners and systems, as provided by the Blizzard Organising Committee are shown in figure 5. The CSTR-Cereproc entry is represented by letter *K*. Given data sparsity problems caused by using a diphone approach, and the lack of any Mandarin language ability in the team, we felt the system performed surprisingly well. It is also important to take into account that the database was collected for syllable coverage rather than diphone coverage. If the data had been collected with diphone coverage in mind then the system would have suffered from less sparsity and performed more effectively.

6. Discussion

6.1. English

Although our English voice was not developed with the intended purpose of consistent neutral speech, but rather focused on realising coherent prosodic styles for genre specific sentences, naturalness and intelligibility tests as evaluated in the Blizzard Challenge

provides valuable insights to the general quality of our voice.

6.2. Mandarin

The main aim of our Mandarin entry was to find out if it was feasible to implement an average-quality Mandarin voice in the CereVoice system in a very limited time (equal to about 10 working days of one person). Given these strict time constraints we focused on the diphone coverage problem only since we believed that it was the highest priority problem. As showed by the results on the Blizzard Challenge, that was sufficient to achieve an average speech naturalness and intelligibility. Additional qualitative data we collected thanks a handful of Chinese native speakers that carried out some short listening (and diagnostic) tests, pointed out the need of a prosodic phrasing module as the next step to further improve the quality of the CereVoice Mandarin voice.

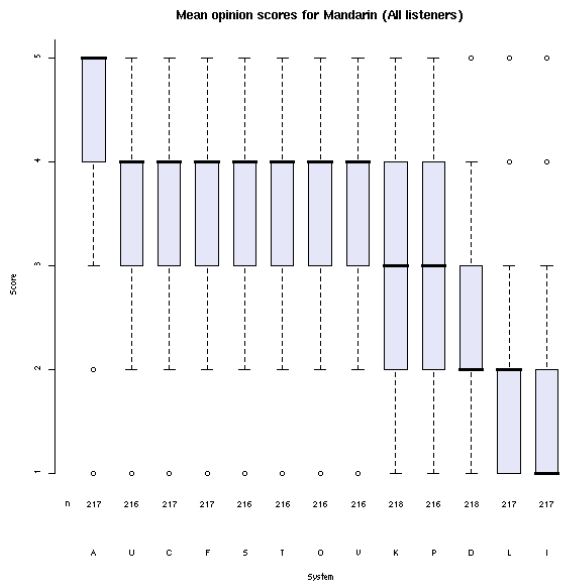

Figure 5: *MOS of the Mandarin voices*

7. Conclusion

The concatenative synthesis approach faces two main challenges. 1. Data sparsity, 2. Joining non homogeneous speech units. A common response to both problems is to reduce the variation within the source database, and collect a database to minimise the prospect of sparsity. However, the damaging effect of data sparsity can be considerably reduced by using appropriate backoff, and in particular, digital signal modification of the speech to shadow missing units. The problems of non-homogeneous data can also be solved by introducing cross genre constraints which carefully prevent joins between incompatible units.

These approaches have to be implemented in a modern concatenative system because of the increasing requirement of expressive and emotional speech synthesis. Such functionality, previously regarded as the domain of research, are increasingly required in day to day commercial applications.

The Tom Sawyer samples at: `http://homepages.inf.ed.ac.uk/janderss/blizzard2008_examples/` exemplifies very well the tension between stability and expressivness that we addressed in this years Blizzard Challenge.

8. Acknowledgements

The CSTR-Cereproc team would like to pay a special thanks to Volker Strom for hands-on assistance and feedback, Chris Pidcock and Cereproc for technical support, and our invaluable Mandarin listeners (Dong Wang, Zhenhua Ling, Zhang Le, Songfang Huang)

9. References

[1] Min Chu, Hu Peng, and Eric Chang, "A concatenative mandarin tts system without prosody model and prosody modification," in *Interspeech*, Perthshire, Scotland, 2001.

[2] Raul Fernandez, Wei Zhang, Ellen Eide, Raimo Bakis, Wael Hamza, Yi Liu, John F. Picheny, Michael Pitrelli, Yong Qing, Zhi Wei Shuang, and Li Qin Shen, "Toward multiple-language tts: Experiment in english and mandarin," in *Interspeech*, Lisbon, Portugal, 2005.

[3] Ren-Hua Wang, Zhongke Ma, Wei Li, and Donglai Zhu, "A corpus-based chinese speech synthesis with contextual dependent unit selection," in *ICSLP*, Beijing, China, 2000.

[4] Volker Strom, Robert Clark, and Simon King, "Expressive prosody for unit-selection speech synthesis," in *Interspeech*, Pittsburgh, U.S.A, 2006.

[5] Volker Strom, Ani Nenkova, Robert Clark, Yolanda Vazquez-Alvarez, Jason Brenier, Simon King, and Dan Jurafsky, "Modelling prominence and emphasis improves unit-selection synthesis," in *Interspeech*, Antwerp, Belgium, 2007.

[6] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X. Liu, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, *The HTK Book*, Cambridge University Engineering Department, 2006.

[7] C. Bennet and A. Black, "The Blizzard Challenge 2006," in *BLZ2-2006*, 2006.

[8] R. Clark, M. Podsiadlo, M. Fraser, C. Mayo, and S. King, "Statistical analysis of the Blizzard Challenge 2007 listening test results," in *BLZ3-2007*, Bonn, Germany, 2007.

[9] Gregor O. Hofer, Korin Richmond, and Robert A.J. Clark, "Informed blending of databases for emotional speech synthesis," in *Interspeech*, 2005.

[10] Marc Schröder, "Expressing degree of activation in synthetic speech," *IEEE Trans. on Speech, Audio, and Language*, vol. Vol. 14, no. No. 4, 2006.

[11] Takashi Nose, Junichi Yamagishi, and Takao Kobayashi, "A style control technique for speech synthesis using multiple regression HSMM," in *ICSLP*, Pittsburgh, U.S.A, 2006, pp. 1324–1327.

[12] Mark Twain, "The Adventures of Tom Sawyer," http://www.gutenberg.org/files/74/74.txt, July 2008.

[13] Robert A.J. Clark, Korin Richmond, and Simon King, "Open-domain unit selection for the festival speech synthesis system," *Speech Communication*, vol. 49, no. 4, pp. 317–330, 2007.