The CereVoice Characterful Speech Synthesiser SDK

Matthew P. Aylett [1] and **Christopher J. Pidcock** [2]

Abstract. CereProc®Ltd. have recently released a beta version of a commercial unit selection synthesiser featuring XML control of speech style. The system is freely available for academic use and allows fine control of the rendered speech as well as full timings to interface with avatars and other animation.

With reference to this system we will discuss current state-of-the-art commercial expressive synthesis, and argue that underlying current approaches to sythesis, and current commercial pressures, make it difficult for many systems to create characterful synthesis. We will present how CereProc's approach differs from the industry standard and how we have attempted to maintain and increase the characterfullness of CereVoice's output.

We will outline the expressive synthesis markup that is supported by the system, how these are expressed in underlying digital signal processing and selection tags. Finally we will present the concept of second pass synthesis where cues can be manually tweaked to allow direct control of intonation style.

1 INTRODUCTION

CereVoice®is a unit selection speech synthesis software development kit (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. The system is designed with an open architecture, has a footprint of approximately 70Mb for a 16Khz voice and runs at approximately 10 channels realtime. The system is a diphone based unit selection system with pre-pruning and a Viterbi search for selecting candidates from the database similar to systems described in [3, 1, 4].

Speech synthesis has progressed enormously since the trademark Stephen Hawking voice which was based on synthesis developed in the mid-eighties. Current systems are acceptable for reading neutral material such as bank balances but sound unacceptable if you use them to read longer texts or more personal information.

We believe this is caused by current approaches to voice building. Most state-of-the-art synthesisers use unit selection to synthesise speech. This approach is based on recording a large database of speech and concatenating small sections of speech together to create new utterances.

The process for recording the database is time consuming (20-30 hours of studio time) and resource intensive. Thus, for commercial systems, a strong focus is made on creating neutral multiple-use voices. In addition, in order to improve concatenation there is an emphasis on reducing the variance of the speech within the database leading, for example, to requesting the source speaker to alter their natural speaking style to make it unnaturally neutral.

This results in voices which are completely inappropriate for expressive characters.

This leads to a vicious circle: commercial synthesis companies don't produce expressive voices so commercial customers can't develop systems using expressive voices. In turn, this forms the perception that there is no market for expressive voices and thus commercial synthesis companies don't create them.

2 EXPRESSIVE SYNTHESIS: Breaking the Deadlock

Four key elements are required for breaking the vicious circle of dull speech synthesis:

1. Voice building must be made more efficient.
   If it becomes possible to build a voice with 10 hours or 6 hours of studio time the incentive for building more voices and making them more expressive is greater. In addition it becomes possible to record a wider variety of speech styles while maintaining a sufficient commercial standard.
2. Control of speech style
   In order to make use of the variation recorded in the voice, it needs to be categorised, or automatically coded, when the voice is built, and the system needs to be able to select material based on this categorisation during synthesis.
3. Semi-automatic synthesis
   Although we don't yet understand how to completely control expressive voices we can use a limited amount of manual intervention to create expressive and characterful cues and prompts. Inserting automatic synthesis between these stock phrases is a pragmatic way of generating expressive dynamic synthesis.
4. Development of applications which require characterful synthesis
   In order to move the technology forwards we need pressure from innovative application developers who can see and harness the enormous potential of characterful synthesis.

CereProc has addressed the first issue by developing a completely automatic voice generation and capture system. This has made the general voice building process more efficient and allows more risks to be taken in the generation of expressive voices. For example a George Bush voice was successfully developed completely automatically from web based material.

In addition CereProc reduces the amount of material required for sound coverage using a process we term 'voice bulking' where unusual diphones (the basic unit used in the synthesis) can be synthetically generated offline. This allows more material to be recorded for prosodic and speech style coverage.

The ability to select and mimic speech styles is accomplished with the use of a rich XML control language. A special tag within this control language also allows the manual manipulation of the synthesis

[1] CereProc Ltd. and CSTR, University of Edinburgh, email: matthewa@inf.ed.ac.uk

[2] CereProc Ltd.

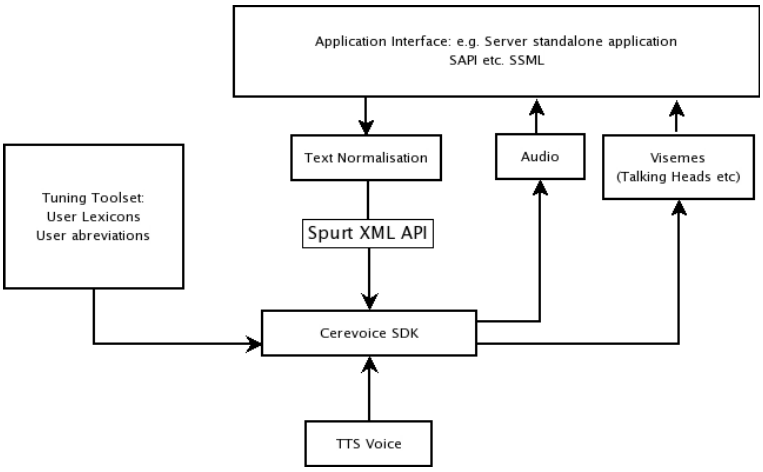

Figure 1. *Overview of the architecture of the CereVoice synthesis system. A key element in the architecture is the separation of text normalisation from the selection part of the system and the use of an XML API.*

process by allowing the user to cycle through the selection of sounds made for a particular word. This allows a simple manual method for discarding the units selected for a word and selecting an alternative set. In many cases this simple operation of discarding unwanted synthesis is sufficient for selecting synthesis which the user finds more appropriate.

Finally, by making the system freely available to the academic community as well as allowing innovative commercial enterprises to take part in an extensive beta test program, CereProc hopes that application developers will make use of this functionality and in turn drive the technology forward.

Despite this, perhaps the most important aspect of creating characterful voices is the simple intention of doing so. In many systems variation in speech style is removed in order to make smoother concatenation easier. CereProc, in contrast, prefers to retain the variation and put more effort in to developing the concatenation process. We have also found that users will accept minor concatenation errors if the voice has more personality. Given that many commercial voices have very few concatenation errors but have a speech style so dull and repetitive that extended synthesis becomes unacceptable, CereProc has found that commercial leverage can be gained by trying to offer voices which sound more characterful and give a stronger impression of a personality behind the voice.

2.1 Overview of the System

CereVoice is a new 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 Figure 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 CONTROLLING EXPRESSIVE SPEECH IN CEREVOICE

The CereVoice front end takes text and generates a series of XML objects we term spurts. The spurt is a section of speech surrounded by pauses. XML markup can be inserted into the input speech and is maintained in the spurt output. The CereVoice system allows a very wide variety of XML markup to control synthesis. Industry standard SSML markup [6] is converted by the front end into a 'reduced instruction set' of XML with a clear functional specification.

In addition, a set of XML markup is allowed which can change the selection process in the system, for example the ability to alter pitch targets. Tags used to alter selection are used in conjunction with tags which cause a change in the speech using digital signal processing to create different speech styles.

The speech styles are based on the activation-evaluation (AE) space, Figure 2. Here emotional states are described in terms of a value varying from very active to very passive and a value varying from very positive to very negative. Within CereVoice 1.2 (alpha) the perception of the emotional content of the speech in terms of the AE space is controlled by four speech style tags: happy (active/positive), calm (passive/positive), cross (active/negative), sad (passive/negative)[3].

Each tag gives a perception of emotion fairly central to each quarter of the AE space. Variation across the positive/negative plane is created by recording two extra sub-sets of data from the speaker. In the first the speaker is requested to produce speech with an unusually relaxed voice quality and in the second set with an unusually tense voice quality. The extent speakers are able to modify their speech in this way, and its relationship to their normal speaking style, varies. This in turn can affect how strongly a change is perceived when the tags are applied. For example, CereProc's Scottish voice 'Heather'

[3] Subject to UK patent application: 0704205.4

was chosen specifically for her cheerful and relaxed speaking style. For this reason the movement into the positive side of the AE space for this voice is less marked than towards the negative side of the space.

Variation across the active/passive plane is achieved using digital signal processing. In general a higher average pitch, a slightly faster speech rate, and increased speech volume make the speech sound more active and whereas the converse make the speech sound more passive.

The intensity of the perceived emotion across the active/passive dimension can thus be altered by changing the underlying control tags that make up the speech style. However their are severe limitations to the extent this is effective. Only a certain degree of modification can be carried out on the speech before it begins to sound unnatural rather than more active or more passive. Thus it is not possible to generate hyper-emotional such as fury, bliss, despair. In contrast, if the changes are too small the change a change in speech style is not perceived at all.

Despite the difficulty of subtle control and the inability to reach edges of the AE space, the use of the tags can be very effective. Much work in altering the perceived emotion of synthetic speech generated using the unit selection approach has concentrated on comparing identical sentences with differences in pitch, duration, rate and voice quality. This is because the content of the sentence has a strong effect on a subjects perception of the emotion in the speech. For a scientific evaluation of the importance of the different cues for the perception of emotion the effect of content is a confounding factor. Fortunately, as a pragmatic engineering solution for adding emotion to a voice, it acts a strong reinforcement to the underlying effects of the speech tags.

This reinforcing effect can be further improved if the negative/positive voice subsets also focus on covering negative and positive vocabulary items.

Overall, the positive/negative voice quality data, the ability to effect unit selection based on pitch and duration features, and the application of rate, pitch and duration changes using digital signal processing act a little like an artists pallet. Creating satisfying emotional characteristics using this functionality is still extremely difficult, just as being able to paint a picture is difficult no matter how many expensive brushes and paints you may have. Making this functionality available in a state of the art commercial synthesiser is, however, a critical step in making characterful synthesis possible.

4 SECOND PASS SYNTHESIS

The vast proportion of speech audio currently used in computer applications is in the form of recorded prompts. This alone demonstrates that although fully automated synthesis is required for completely dynamic content, much content is, in fact, not that dynamic at all. Currently, users of speech synthesis have used markup such as SSML [6] to manually control exactly how synthesis is realised. However the format of much of this markup stems from earlier diphone based synthesis systems rather than database approaches. CereVoice, however, will accept markup which allows users to control the inner working of the selection process. Such manual intervention is an effective stop-gap technique for competing with natural pre-recorded prompts.

Second-pass synthesis is a post-hoc method of tuning the synthesis output to improve the perceived quality of the output. A Viterbi search is used to find the ’best’ sequence of states. In CereVoice it is possible to ask the engine to prune out a section of the best path found during the Viterbi search and to rerun the Viterbi over that section to find a less optimal alternative or *variant*. The next variant approach can be applied to a whole utterance or, more usefully, focus on a problem word or diphone. In the case of changing a single word or diphone in a larger utterance, units not within the the variant section are ’locked’ to prevent modification of units that are considered acceptable. A new variant is selected by running the Viterbi search then pruning out the rejected selection of units. The pruning out of rejected units is cyclical, continuing until the requested variant number is found. Inside an XML spurt, a word can be enclosed by a ’usel’ tag containing a variant attribute to force this behaviour. For example <usel variant=’0’> is equivalent to no tag, and <usel variant=’6’> would be the sixth alternative according to the Viterbi search. Fig. 3 shows a schematic of this process.

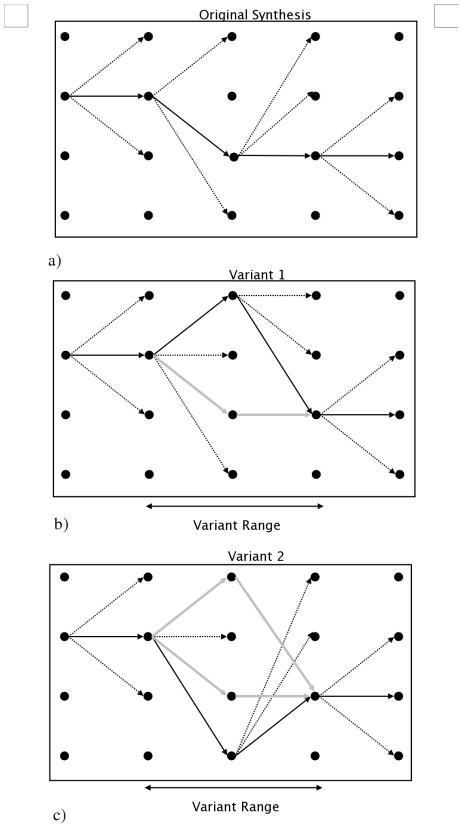

Figure 3. *Schematic diagram of the CereVoice variant tag process. a) The best path chosen by the Viterbi is shown as a black line. b) The unit in row 3 column 3 is rejected and the variant tag requests the next alternative. The path going through the unit is pruned out and a second path marked in black is selected. c) The new unit at row 1 column 3 is also rejected, the process is run again, a final acceptable unit at row 4 column 3 is selected.*

Below is an example of text marked up with variant tags.

```
The <usel variant=’2’>Fruitto</usel> de
Mare featured, calamari served with <usel
variant=’1’>tomatoes</usel>, peppers,
artichoke, avocado and, again, frisee.
```

Investigating efficient manual methods for improving synthesis addresses a crucial research question; given the database, how good could the synthesis become if our search algorithms produced optimum quality speech? In order to supply synthesis for entertainment there is a requirement for building fast, good quality characterful voices, often within specific domains. It is currently unclear what the degrees of freedom are for minimising the size of col-

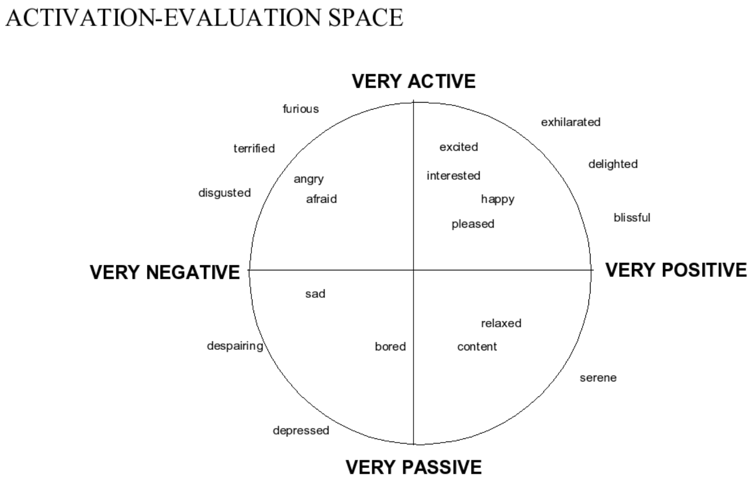

Figure 2. *Activation-Evaluation space. Adapted from [5] in turn adapted from [2]*

lected databases. Previous work which has tried to improve the quality of voices made from small databases has made use of information from a different voice with a larger database, either by using voice-morphing e.g.[3] or the larger voices prosodic model e.g.[4]. In contrast, second-pass synthesis allows us to answer the question of whether critical errors in the synthesis are caused by the poverty of the search algorithm or whether they are caused by database sparsity.

5 CASE STUDIES

In order to demonstrate the use of the XML control language we will present two case studies which show how they can be used. The first is an example of how the underlying tags in our Scottish voice are used to position the speech within the AE space for the 'happy' tag. The second is how we can use manual intervention to tailor a short paragraph of speech synthesised using our George Bush voice. Examples of the audio for these two case studies are available at http://www.cogsci.ed.ac.uk/∼matthewa/AISB2007.html.

5.1 Case Study 1: The Happy Tag

In order to explore how we create our happy speech style tag we will start by synthesising material which should be spoken happily in this example the sentence 'What a lovely day.' As discussed earlier, attempting to alter the emotional bias of the content is extremely difficult and will not be attempted here.

The baseline for this sentence is synthesised with the raw text:

```
What a lovely day.
```

The first stage in the process is to bias the unit selection to choose units from the calm voice quality section of the database. This is accomplished using a *genre* attribute within the unit selection tag *usel*.

```
<usel genre=′calm′>What a lovely day.</usel>
```

This makes a major impact on the material selected and immediately produces a more positive sounding utterance. It sounds cheerful but not as upbeat as we might like. In order to make it sound more active we can in turn: increase the average pitch by 5 hertz,

```
<usel genre=′calm′><sig f0=′+5′>What a
lovely day.</sig></usel>
```

increase the amplitude. The value '2.0' used here does not directly increase the amplitude by two times its original value. In order to prevent clipping the speech is also compressed so that higher volume sections are not amplified as much as quiet sections.

```
<usel genre=′calm′><sig f0=′+5′
amplitude=′2.0′>What a lovely
day.</sig></usel>
```

and increase the speech rate.

```
<usel genre=′calm′><sig f0=′+5′
amplitude=′2.0′ rate=′1.05′>What a lovely
day.</sig></usel>
```

The combined effect is quite subtle but reasonably effective. The effects of the digital signal processing are more pronounced if you compare it do doing the opposite with the speech, i.e. reducing the pitch. lowering the amplitude and slowing the speech rate. The effect of this is to produce a stronger feeling of calm.

```
<usel genre=′calm′><sig f0=′-5′
amplitude=′0.5′ rate=′0.95′>What a lovely
day.</sig></usel>
```

it is not possible to use digital processing techniques to make increase the percept of happiness much more than this. For example if we continue to increase pitch, amplitude and rate it begins to sound strange.

```
<usel genre=′calm′><sig f0=′+15′
amplitude=′3.0′ rate=′1.2′>What a lovely
day.</sig></usel>
```

In our commercial system these underlying control tags are bundled into a SSML style tag <voice emotion='happy>.[4]

5.2 Case Study 2: George Bush Discusses HRI

The CereProc George Bush voice was created using audio trawled from the web. Unlike CereProc voices, where the design and capture of the audio is within our control, there is no guarantee of having appropriate coverage of phonetic material or that the acoustics will

[4] In our current system we use a lower amplitude increase in this bundled tag because we are currently unhappy with audible artifacts at the higher level described here.

be at the same standard we expect from a bespoke recording environment. In addition the transcriptions lifted from the web can be slightly inaccurate and that can cause quite serious synthesis errors.

For this reason the George Bush voice offers an excellent example of how we can remove mistakes and improve synthesis with a little manual intervention.

The text we chose to synthesise was taken from the first two sentences of the description of scope of the special session in AISB on language, speech and gesture for expressive characters.

```
Research into expressive characters, for
example embodied conversational agents, is a
growing field, while new work in human-robot
interaction (HRI) has also focused on
issues of expressive behaviour. With recent
developments in computer graphics, natural
language engineering and speech processing,
much of the technological platform for
expressive characters  both graphical and
robotic  is in place.
```

The raw synthesis of this material using the George Bush voice was reasonably acceptable but did contain some errors. Below is a marked up version of the text which gives a better rendition. The superscript beside each tag links to an explanation for its insertion below.

`Research` **`<lex phonemes='ih2 n t uw1'>`** `into`[1] **`</lex>`** **`<usel variant='1'>`** `expressive`[2] **`</usel>`** `characters, for example embodied conversational agents, is a growing field,` **`<break type='4'/>`**[3] `while new work in` **`<sig rate='0.8'>`** `human-robot`[4] **`</sig>`** **`<lex phonemes='ih1 n t er0 ae1 k sh ax0 n'>`** `interaction`[5] **`</lex>`** **`<break type='0'/>`**[6] `(HR` **`<usel variant='3'>`** `I`[7] **`</usel>`**`) has also focused on issues of expressive` **`<lex phonemes='b ax0 hh ey1 v y er0'>`** `behaviour`[8] **`</lex>`**`. With recent developments in computer graphics, natural language engineering and speech processing,` **`<break type='4'/>`**[9] `much of the technological` **`<usel variant='1'>`** `platform`[10] **`</usel>`** `for expressive characters both graphical and robotic is in` **`<usel variant='2'>`** `place`[11] **`</usel>`**`.`

The explanations for the additional tags are as follows:

1. The default stress on 'into' is to reduce it (i.e. 'inter' rather than 'intoo'). We override the pronunciation and thus the reduction with this tag.
2. There is a error caused by the database which produces something which sounds more like 'ixpressive' than 'expressive'. The variant tag discards this selection and the next selection does not have the error.
3. A comma normally generates an intermediate phrase break. In this case a the more final break '4' is appropriate. (Replacing the comma with a full stop would have had the same effect).
4. 'human-robot' is an unusual compound. A human speaker would typically make this more salient and the same effect can be achieved by using digital signal processing to slow the speech rate down by 20%.
5. It is hard to select the correct stress of syllables like 'in' in 'interaction'. By using the phoneme tag we have increased the stress from the default of secondary to primary by adding a '1' on the phone 'ih'.
6. The bracket creates a non-final phrase break by default. This has been removed by using a break of type '0' which prevents an odd pause before the acronym.
7. getting the stress right in acronyms is difficult. We want the voice to say hc**I** not h**C**i. We reject the first 0-2 variants of 'I'for being too reduced and use the variant '3' version.
8. The voice is a general American voice and doesn't have a lexical entries for British spellings. This is the US pronunciation of the word 'behavior'.
9. See note 3.
10. Again getting the stress right on compounds is difficult. We preferred the stress on the variant '1' to the original.
11. George bush doesn't have very much phrase final intonation in his speeches. Like many politicians he has learnt the trick of not sounding finished as he talks. Variant '2' was the first variant with a satisfying phrase final intonation.

This may seem a lot of manual work to get your synthesis to sound better. However, bear in mind we are using a voice that is not designed for this sort of synthesis. Most of the changes are actually using appropriate phrasing (spoken language has shorter sentences than written language), ensuring pronunciation is correct and fixing the odd concatenation error with a variant tag.

In this case, its also worth bearing in mind that getting George Bush into the recording studio and get him to say it perfectly is intractable, and even with more accessible voice talents re-recording material is a resource intensive and troublesome job.

Even if voices are constructed from limited prompt material, as the original prompts will be generated perfectly, we believe it is almost foolish not to use a synthesis solution to allow greater flexibility. After all, it offers more control and the possibility of creating new material without having to re-record.

6 CONCLUSION

Speech synthesis is a key enabling technology for pervasive computing. For many areas a key requirement is that the user is communicating with something which can simulate character and personality. Much current speech synthesis, although of a high standard for generating neutral speech, falls far short of what is required for giving character to avatars and speech based systems. Although there is much we do not understand in the generation of expressive speech it is possible to generate limited expressive speech and to further increase its effectiveness by offering more manual control of the speech rendered when required.

By making this technology freely available to the research establishment we hope to increase the awareness of this functionality, improve it and discover the extent it can produce innovative applications and user experiences.

REFERENCES

[1] Robert A.J. Clark, Korin Richmond, and Simon King, 'Festival 2 - build your own general purpose unit selection speech synthesiser', in *5th ESCA Workshop in Speech Synthesis*, pp. 147–151, (2004).

[2] Roddy Cowie, Ellen Douglas-Cowie, Susie Savvidou, Edelle McMahon, Martin Sawey, and Marc Schröder, 'Is disfluency just difficult?', in *ISCA Tutorial and Research Workshop on Speech and Emotion*, (2000).

[3] A.J. Hunt and A.W. Black, 'Unit selection in concatanative speech synthesis using a large speech database', in *ICASSP*, volume 1, pp. 192–252, (1996).

[4] John Kominek, Christine L. Bennet, Brian Langer, and Arthur R. Toth, 'The Blizzard challenge 2005 CMU entry - a method for improving speech synthesis systems', in *INTERSPEECH*, pp. 85–88, (2005).

[5] R. Plutchik, *The Psychology and Biology of Emotion*, Harper Collins, New York, 1994.

[6] Paul Taylor and Amy Isard, 'SSML: A speech synthesis markup language', *Speech Communication*, **21**, 123–133, (1997).