EXPROS: A Toolkit for Exploratory Experimentation with Prosody in Customized Diphone Voices

Joakim Gustafson and Jens Edlund

KTH Speech Music and Hearing
{jocke,edlund}@speech.kth.se

Abstract. This paper presents a toolkit for experimentation with prosody in diphone voices. Prosodic features play an important role for aspects of human-human spoken dialogue that are largely unexploited in current spoken dialogue systems. The toolkit contains tools for recording utterances for a number of purposes. Examples include extraction of prosodic features such as pitch, intensity and duration for transplantation onto synthetic utterances, and creation of purpose-built customized MBROLA mini-voices.

1 Introduction

Prosodic features, such as pitch, intensity and duration, play an important role for many of the aspects of spoken dialogue that are prolific in and central to human-human dialogue, yet to date rarely exploited in human-computer dialogues. Examples include interaction control, the management of turn-taking, interruptions, and back-channels; attitude towards what is said, such as the signalling of uncertainty or certainty; prominence, such as contrastive focus and stress; and grounding, as in brief feedback utterances for verification and clarification. There is a fair body of research into these matters from the spoken dialogue system point of view on the perception side, and some of which has been taken as far as to implementation and experimentation in full-blown spoken dialogue systems. On the production side, there are fewer examples where our knowledge of prosody has made it all the way to full-blown systems. In current spoken dialogue systems, pre-recorded prompts or unit selection synthesis are often chosen for the voice quality. The drawback is that these techniques make it difficult to vary prosody and to control this variation in any detail, so few examples of experimentation with such variations exist (see [Raux and Black, 2003], however, for an example and an overview]. Although there is a large body of studies of prosodic features using re-synthesis with modified prosody (using e.g. Praat[1]) and with HMM synthesis, the results have proven difficult to implement in real on-line systems. Other synthesis methods providing greater control over prosodic features are formant synthesis and diphone synthesis. The relatively low voice quality of formant synthesis makes it unsuitable for many user studies, however, and diphone synthesis suffers from the relatively large cost of recording the required diphones, as well as from less-than-perfect voice quality.

[1] http://www.fon.hum.uva.nl/praat/

E. André et al. (Eds.): PIT 2008, LNAI 5078, pp. 293–296, 2008.

This paper presents EXPROS, a graphical toolkit that allows us to experiment with prosodic variation in diphone synthesis in a more efficient manner. Before going into the functionality currently built into the toolkit, lets discuss a few of its applications. Our main reason to experiment with prosodic variation is to make spoken dialogue systems that more closely mimic human-human dialogue, in order to better exploit its strengths. This need not be the case for all spoken dialogue system design, but it is our motivation here. The following are three examples of increasing complexity of dialogue needs that EXPROS aim to meet. Each of the examples is illustrated in the demonstration.

(1) A key area where humans excel over current spoken dialogue systems is interaction control, the management of the flow of the dialogue, for example turn-taking and interruptions. An oft-mentioned problem is that of user barge-ins, but we would also want our systems to be able to deal with system barge-ins and self-interrupts in a better manner. The following dialogue excerpts exemplify this:

```
U   What's the weather like in Stockton?
S   The weather in Stockholm? Wait a mo* [*ment, I'll look it up]
U                                      No, I said Stockton

U   Any news on fashion /SIL/ in Tibet?
S                       OK, le* [*t me see what I can do]
S                                Ah, let me see what I can do

U   Are there any news about Camden market?
S   Let me see... no, there's no* [*thing new at the moment]
                             /fresh news arrive/
S                                 Oh, hang on, there's a fire in Camden!
```

In order for a spoken dialogue system to produce the behaviours listed above, the system's processing in its entirety needs to be incremental, as noted in [Allen et al., 2001] and [Aist et al., 2006]. Here, however, we are only concerned with being able to control the rendering of the speech sounds sufficiently to produce utterances like the ones above.

(2) In order to achieve this kind of dialogue, we need to be able to test variations in perception tests as well as in real human-computer dialogue situations. To do this, we need to be able to record the required prompts with different prosody, at the very least. In many cases, we may want to record new diphones – in the example above, for example, we could record P*_SIL diphones, that go from a phoneme P to silence SIL abruptly, to make the interruptions sound more realistic. Recording extra sets of diphones for hypo- and hyper-articulated speech may also be useful, as well as affective speech, for example angry or despondent. Testing out new voices can be very time consuming, however, as a Swedish diphone voice typically contains some 5000 diphones. This is far too expensive for exploratory studies into the effects of prosodic and voice quality variations. Instead we can create mini-voices – voices with few diphones, that are able to produce only a limited number of utterances, but that are easy to record and to modify.

(3) Finally, pre-recorded prompts, unit selection synthesis, and diphone synthesis all suffer from the need to enrol the original speaker each time the voice is to be extended or changed. A diphone voice production is furthermore often created in one go, and rarely updated or changed after its completion. We attempt to make it possible

for speakers who are not the original speaker to do as many extensions as possible – particularly to record new prosodic patterns, and also for the voice creation to be done incrementally, by making it simple to add new diphones and diphone sets *when they are needed*.

Prompts and voices developed in EXPROS can be used in perception tests, either of standalone prompts or of re-synthesised dialogue utterances, but most importantly they are intended for use in interactive experiments, where the pragmatics – the actual effect prosodic variation has on the interaction – can be measured.

2 The EXPROS Toolkit

The toolkit uses the Snack sound toolkit[2] as its backbone, and integrates functions from a number of existing tools, such as the Mbrola engine and database builder[3], a PC-KIMMO[4] morphological dictionary, NALIGN forced alignment [Sjölander and Heldner, 2005], **`/nailon/`** prosodic extraction and normalisation [Edlund and Heldner, 2005], etc.

Text processing: Reading and management of (prosodic) labels in the orthographic input. These labels could be used to generate prosodic patterns automatically, such as increased stress or prolonged syllables.

Grapheme to phoneme conversion: The toolkit currently incorporates automatic transcription using PC-KIMMO and a Swedish dictionary with transcribed morphs, an NALIGN CART tree built on Centlex, a Swedish pronunciation dictionary developed at the Centre of Speech Technology, as well as a set of cooarticulation rules (over word boundaries) built into NALIGN. In addition, user lexica can be defined and used.

Automatic speech alignment: The toolkit uses the forced aligner NALIGN to extract phone start and end times from recordings.

Automatic prosody parameter extraction: For prosodic analysis, the toolkit can currently use the methods built into the Snack sound toolkit (ESPS get_f0 and AMDF pitch extraction as well as power analysis, which can be used to estimate spectral tilt). The normalization methods built into **`/nailon/`** are also available.

Modification of prosodic parameters: The toolkit provides a number of methods for modification of prosodic parameter curves as well as creation of new curves. These include direct manipulation in a GUI, stylisation, normalisation and transformation to another speakers speaking style, model generated prosodic curves, and transplantation of curves from recordings.

Diphone synthesis: The toolkit uses an extended MBROLA synthesis engine [Drioli et al., 2005] which adds control of for example gain, spectral tilt, shimmer and jitter to render audio. Using a combination of the components listed above, the toolkit also gives the possibility to automatically generate the data needed to build new

[2] http://www.speech.kth.se/snack/
[3] http://tcts.fpms.ac.be/synthesis/mbrola.html
[4] http://www.sil.org/pckimmo/

MBROLA diphone databases, and some scripts to make on-the-fly modifications to how the MBROLA engine select diphones.

3 Conclusions

Preliminary listening tests suggest that transplanting durations, intensity and pitch from human recordings onto the diphone synthesis makes diphone voices sound considerably better as a whole, which is promising. We have for example used the EXPROS tool to improve the subjective ratings of a bad speaker, by re-synthesizing 30 seconds of his speech with increased pitch variation and speaking rate [Strangert and Gustafson, Submitted]. Examples are included in the demonstration.

The toolkit has also proven valuable for verifying the quality of automatic prosodic analysis – pitch and intensity extraction as well as phone durations – by listening to the original recording and its resynthesis in parallel – a method inspired by Malfrere & Dutoit [Malfrere and Dutiot, 1997].

Acknowledgements

Thanks to everyone who has put hard work on developing the publically available tools that are used in this toolkit. Special thanks to Thierry Dutoit and MBROLA and Kåre Sjölander (Snack/NALIGN). This work was supported by the Swedish research council project #2006-2172 (Vad gör tal till samtal/What makes speech special) and MonAMI, an Integrated Project under the EC's Sixth Framework Program (IP-035147).

References

[Aist et al., 2006] Aist, G., Allen, J.F., Campana, E., Galescu, L., Gómez Gallo, C.A., Stoness, S.C., Swift, M., Tanenhaus, M.: Software Architectures for Incremental Understanding of Human Speech. In: Proceedings of Interspeech, Pittsburgh PA, USA (2006)

[Allen et al., 2001] Allen, J.F., Ferguson, G., Stent, A.: An architecture for more realistic conversational systems. In: Proceedings of the 6th international conference on Intelligent user interfaces, pp. 1–8 (2001)

[Drioli et al., 2005] Drioli, C., Tesser, F., Tisato, G., Cosi, P.: Control of voice quality for emotional speech synthesis. In: Proceedings of AISV 2004, Padova, pp. 789–798 (2005)

[Edlund and Heldner, 2005] Edlund, J., Heldner, M.: Exploring prosody in interaction control. Phonetica 62(2-4), 215–226 (2005)

[Malfrere and Dutiot, 1997] Malfrere, F., Dutoit, T.: Speech synthesis for text-to-speech alignment and prosodic feature extraction. In: Circuits and Systems: Proc. of ISCAS (1997)

[Raux and Black, 2003] Raux, A., Black, A.: A Unit Selection Approach to F0 Modeling and its Application to Emphasis. In: Proceedings of ASRU 2003, St Thomas, US (2003)

[Strangert and Gustafson, Submitted] Strangert, E., Gustafson, J.: Subject ratings, acoustic measurements and synthesis of good-speaker characteristics, Interspeech 2008, Brisbane (submitted, 2008)

[Sjölander and Heldner, 2005] Sjölander, K., Heldner, M.: Word level precision of the NALIGN automatic segmentation algorithm. In: Proc of The XVIIth Swedish Phonetics Conference, Fonetik 2004, pp. 116–119. Stockholm University (2004)