EXPROS: Tools for exploratory experimentation with prosody

Joakim Gustafson and Jens Edlund
Centre for Speech Technology, KTH Stockholm, Sweden

Abstract

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

Introduction

This demo paper presents EXPROS, a graphical toolkit permitting us to experiment with prosodic variation in diphone synthesis in an efficient manner.

Prosodic features such as pitch, intensity and duration play an important role for many of the aspects of spoken dialogue that are central to human-human dialogue. Still, to date they are rarely exploited in human-computer dialogues. Examples of areas that would benefit from the inclusion of more prosodic knowledge include interaction control, the management of turn-taking, interruptions, and backchannels; 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.

On the perception side, there is a fair body of research into these matters from the spoken dialogue system point of view. Some of these results have 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 because of their superior 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. One of the few examples is Raux & Black (2003), which also provides an overview of the topic. There is a large body of studies of prosodic features using re-synthesis with modified prosody (using e.g. Praat) and with HMM synthesis, but the results have proven difficult to implement in real on-line systems.

Other synthesis methods – formant synthesis and diphone synthesis – provide greater control over prosodic features. 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.

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.

Interaction control

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 dialogue excerpts in Example 1 exemplify this. 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 in Example 1.

Example 1: Three dialogue excerpts

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

The sounds of dialogue

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.

Incremental development

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 stand-alone 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.

The EXPROS Toolkit

The toolkit uses the Snack sound toolkit[1] as its backbone, and integrates functions from a number of existing tools, such as the Mbrola engine and database builder[2], a PC-KIMMO[3] morphological dictionary, NALIGN forced alignment (Sjölander & Heldner, 2004), `/nailon/` prosodic extraction and normalisation (Edlund & Heldner, 2006), 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

[1] http://www.speech.kth.se/snack/

[2] http://tcts.fpms.ac.be/synthesis/mbrola.html

[3] http://www.sil.org/pckimmo/

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

Next steps

A number of experiments and investigations using EXPROS are underway:

We will test the effects of transplanted prosody on perceived synthesis quality. 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 also want to test this in the context of the findings of Hjalmarsson & Edlund (in press), where synthesised utterances containing typical features of human-human dialogue, such as filled pauses and repetitions, were investigated.

The EXPROS tool has recently been used to improve the subjective ratings of a bad speaker, by re-synthesising 30 seconds of speech with increased pitch variation and speaking rate (Strangert & Gustafson, submitted). We intend to do more experiments with resynthesis in order to explore the limits of what can be expressed by manipulating prosody alone.

Furthermore, the toolkit has 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 (1997).

Finally, we are in the process of running tests were subjects use EXPROS to create new versions of very brief feedback or clarification utterances in order to change their meaning. We have previously shown that monosyllabic words can be understood as positive or negative grounding on the perceptual or understanding levels by manipulating their pitch contour (Edlund et al., 2005), and using EXPROS, we hope to be able to show the same for multisyllabic compound words.

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 (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, G., Allen, J. F., Campana, E., Galescu, L., Gómez Gallo, C. A., Stoness, S. C., Swift, M., & Tanenhaus, M. (2006). Software Architectures for Incremental Understanding of Human Speech. In *Proceedings of Interspeech* (pp. 1922-1925). Pittsburgh PA, USA.

Allen, J. F., Ferguson, G., & Stent, A. (2001). An architecture for more realistic conversational systems. In *Proceedings of the 6th international conference on Intelligent user interfaces* (pp. 1-8).

Drioli, C., Tesser, F., Tisato, G., & Cosi, P. (2005). Control of voice quality for emotional speech synthesis. In *Proceedings of AISV 2004* (pp. 789-798). Padova, Italy.

Edlund, J., & Heldner, M. (2006). /nailon/ - software for online analysis of prosody. In

Proc of Interspeech 2006 ICSLP. Pittsburgh PA, USA.

Edlund, J., House, D., & Skantze, G. (2005). The effects of prosodic features on the interpretation of clarification ellipses. In *Proceedings of Interspeech 2005* (pp. 2389-2392). Lisbon, Portugal.

Hjalmarsson, A., & Edlund, J. (in press). Human-likeness in utterance generation: effects of variability. To be published in *Proceedings of the 4th IEEE Workshop on Perception and Interactive Technologies for Speech-Based Systems*. Kloster Irsee, Germany.

Malfrere, F., & Dutoit, T. (1997). Speech Synthesis for Text-to-Speech Alignment and Prosodic Feature Extraction. In *Speech Synthesis for Text-to-Speech Alignment and Prosodic Feature Extraction", F. Malfr&#65533;re & T. Dutoit, Proceedings of the International Symposium on Circuits and Systems* (pp. 2637-2640,).

Raux, A., & Black, .. (2003). A Unit Selection Approach to F0 Modeling and its Application to Emphasis. In *Proceedings of ASRU 2003, St Thomas, US Virgin Islands.*.

Sjölander, K., & Heldner, M. (2004). Word level precision of the NALIGN automatic segmentation algorithm. In *Proc of The XVIIth Swedish Phonetics Conference, Fonetik 2004* (pp. 116-119). Stockholm University.

Strangert, E., & Gustafson, J. (submitted). Subject ratings, acoustic measurements and synthesis of good-speaker characteristics. Submitted to *Proceedings of Interspeech 2008*. Brisbane, Australia.