Synthesis Units for Conversational Speech
- Using Phrasal Segments -

Nick Campbell

ATR Network Informatics Lab
"Keihanna Science City", Kyoto 619-0223

nick@atr.co.jp

Abstract

This paper describes the use of phrase-sized segments for the concatenative synthesis of conversational speech and discusses the differences in selection criteria that become necessary when the source corpus contains several years of conversational speech samples. It claims that natural-sounding conversational speech can be reproduced by use of such phrase-sized chunks for concatenation, and that their physical adjacency in the speech enhances the sense of continuity in spite of their discrete origins in the corpus.

1. Introduction

The author previously proposed the CHATR system [1,2] of prosody-based unit-selection for concatenative waveform synthesis, and now extends this work to incorporate the results of an analysis of almost five-years of high-quality recordings of spontaneous conversational speech in a wide range of actual daily-life situations. Having such an enormous corpus of speech samples available for concatenative synthesis allows us to consider the selection of complete phrase-sized segments from a discourse, and thereby changes the focus of unit selection from that of segmental or phonetic continuity to one of prosodic and discoursal appropriateness instead. The paper therefore describes the characteristics of conversational speech in the context of corpus-based speech synthesis. Samples of the resulting large-corpus-based conversational synthesis (and an extended version of this paper [3], with a set of powerpoint slides) can be found at http://feast.his.atr.jp/AESOP.

2. Expression of Information and Affect

Previous work [4,5] has already described the differences between expression of affect and information in the JST/ATR ESP corpus. We have proposed that any given conversational speech utterance can be categorised into either I-Type or A-Type classes [6], where I-Type indicates a predominance of propositional content, and A-Type indicates a predominance of affect (or so-called 'Kansei' information) in the utterance.

I-Type utterances tend to be longer, are grammatically rich, and can usually be safely characterised by a transcription of their linguistic content alone. They can often be adequately synthesised by current speech synthesis technology. On the other hand, A-Type utterances *can not be* adequately understood from just a transcription of their linguistic content alone, and they require a detailed prosodic specification in addition, to indicate the significant speaker-state and speaker-listener relationships that are displayed through meaningful variations in the speaking-style and voice quality information.

対話型音声合成：フレーズ単位の利用について
ニック・キャンベル　ATR ネットワーク情報学研究所.

3. A Framework for Defining Speaking-Style

In order to synthesise the A-type utterances, we need to know who is talking to whom (not their names, of course, but their social and inter-personal relationships), and where, and why. An utterance whose primary function is to display affect will be either of a non-lexical type (typically short simple repeated monosyllables, e.g., "yeah-yeah-yeah-yeah-yeah", or "uhuh, uhuhuh") or a common phrase, such as "Hi there, how are you?", which is used more for its phatic function rather than for the transmission of propositional content. These non-verbal and often non-lexical 'grunts' make up as much as half of the utterances in the ESP corpus and can in many cases be reliably 'understood' even across language boundaries [7].

This display of affect as a speech event can be coded in higher-level terms as a combination of the following three features, or 'SOE' constraints: (i) Self, (ii) Other, (iii) Event, as in equation (1) which defines an utterance (U) as (probably uniquely) specified by the realisation of a *discourse event* (E) given context-pair *self* (S) and *other* (O) *note that this differs slightly in form from that presented previously in [8]

$$U = E \mid (S , O) \qquad (1)$$

where the feature Self can take different values (representing strong and weak settings with respect to the dimensions *mood* and *interest* respectively) and the feature Other can also take different values (representing strong and weak settings with respect to the dimensions *friend* and *friendly* respectively (see below)), and the feature Event represents a discourse move or a speech act; i.e., the purpose or function of a given utterance.

The feature *Self* refers to (a) the personal state of the speaker and (b) his or her interest in the content of the utterance. For example, a healthy, happy, person is likely to speak more positively than an unhealthy or a miserable one. One who is interested in a topic and/or highly motivated by the discourse is likely to be more active and expressive than one who is not

The feature *Other* refers both to (a) the relationships between speaker and hearer, and (b) the constraints imposed by the discourse context. A person talking with a friend is likely to be more casual or relaxed than one talking with a stranger, but will also probably be more relaxed when talking informally, e.g., in a pub, than when talking formally, e.g., in a lecture hall.

The constraint framework defined by given settings of the SO parameters has a controlling influence on both the content and the expressivity of the utterance that instantiates an event, and has effects on the text as well as the style of the utterance. Because of the great variety in forms of essentially similar utterances, we propose a text-free specification for the input

Table 1. Unit-selection for A-TypeUtterances

CLASS: …….. speech & discourse act/event
constrain the lexical choice of an utterance
VARIANT: …. mood, emotion \& politeness
filter among the utterance-level candidates
TOKENS: …. overall vocal/prosodic settings
select the best by using a continuity filter
SEGMENT: … the speech waveform for output
replay the entire phrase-segment waveform

4. Phrasal Segments as Synthesis Units

With a small speech database, the task of unit selection is to define a measure which minimises two costs simultaneously; i.e., a *target cost* (for prosodic and spectral appropriateness) and a *join cost* (for smoothness of concatenation between the segments) [9]. However, when the source database includes e.g., 5 years of daily conversational speech, as ours does now [10], then the needs for unit selection change drastically.

With a very large source corpus, there is no longer a need to predict or select segments according to sentence-internal prosodic characteristics, since sentence-sized chunks can often be taken whole from the corpus. They will of course sound natural because they are natural; no 'synthesis' is involved. However, when re-using complete utterances or phrase-sized segments from a large corpus, there is a need not just to maintain (join-cost) continuity throughout the discourse, but also to control the affective `colouration` of the utterances appropriately according to the (target-cost) pragmatic needs of the dialogue. That is, we now need to discriminate between sentences or phrases which have the same or similar text content, but which have been spoken in different speaking-styles to display different affective content.

It is of course still necessary to match the overall acoustic characteristics of a given utterance or phrase to those of the previous and following utterances or phrases, so that the output speech does not appear to come from different speakers (as it might if segments from two completely different utterance contexts were selected for contiguous replay) but since phrase-sized utterances can be extracted whole from the corpus, there is no longer any need to model the phrase-internal linguistic prosodic characteristics. The 'target-cost' of unit-selection can be replaced by higher-level selection constraints at the prosody-based filtering stage.

5. Flow of Processing for Unit Selection

Table 1 illustrates the flow of processing for unit-selection. The precise wording of the Event is preferably undetermined in the input, leaving more freedom for the selection of the most appropriate utterance which matches the combined set of selection constraints. These can be hierarchically organized into CLASS (greet, confirm, complain, laugh, accept, decline, etc.,) and VARIANT (happy, sulky, warm, friendly, relaxed, distant, etc.,) in order to determine the initial set of candidate TOKENS from the corpus, from which we choose an optimal SEGMENT according to the above-mentioned continuity constraints. The phrasal segments sent to the audio device, although actually unrelated and disjointed, appear to take on a continuity that is conversationally natural in the resulting dialogue speech. (examples of such synthesized conversations can be heard at the web-page noted above [3b])

6. Conclusion

We have found that as the source speech corpus increases in size and naturalness, so the speech synthesis process moves from the reproduction of phonetic sound sequences for the representation of linguistic information, to the reproduction of speaking styles and voice qualities for the expression of discourse-related and interpersonal affective content. In parallel with this progression, we see that the role of prosodic information has evolved from the simple task of marking boundaries and focal-points to the more complex one of displaying fine details of speaker state and speaker-listener relationships. By re-using small phrase-sized chunks of speech, conversational turns can be reproduced with very high naturalness [5], but the task of filtering these chunks according to higher-level discourse-related constraints requires some knowledge of the interpersonal and affect-related information which cannot be derived from a text alone. The conversational speech synthesizer will require an input modality that allows specification of these higher-level relationships, perhaps *instead of* a specification of the exact text to be spoken.

Acknowledgements

This work is supported partly by a grant from the Japan Science & Technology Corporation and partly by the National Institute of Information and Communications Technology (NiCT). The author is grateful to the management of ATR for their continuing encouragement and support.

References

[1] Campbell, W.N., "Synthesis units for natural English speech'", Transactions of the Institute of Electronics, Information and Communication Eng, SP 91-129, 55-62, 1992.
[2] Campbell, W.N., "CHATR: A High-Definition Speech Re-Sequencing System", in proc Eurospeech'95, Madrid, 1995.
[3] Campbell, W. N., "Developments in corpus-based speech synthesis: Approaching natural conversational speech", in IEICE Transactions, Special Issue on Corpus-based Speech Technology (forthcoming). 2004.
[3b] http://feast.his.atr.jp/AESOP --- synthesis samples
[4] Campbell, W. N., "Listening between the lines; a study of paralinguistic information carried by tone-of-voice", pp 13-16 in Proc International Symposium on Tonal Aspects of Languages, TAL2004, Beijing, China, 2004.
[5] The JST/CREST Expressive Speech Processing project, introductory web pages at: http://feast.his.atr.jp
[6] Campbell, W. N., "Extra-Semantic Protocols; Input requirements for the synthesis of dialogue speech", pp.221-228 in Andre E., Dybkjaer, L., Minker, W., & Heisterkamp, P., (Eds) *Affective Dialogue Systems*, Springer Lecture Notes in Artificial Intelligence Series, 2004.
[7] Campbell, N., & Erickson, D., "What do people hear? A study of the perception of non-verbal affective information in conversational speech", pp. 9-28 in Journal of the Phonetic Society of Japan, Vol 7, Num 4, 2004.
[8] Campbell, W. N., "User Interface for an expressive speech synthesiser", 1-7-21 in Proc Spring mtg of the ASJ, 2004.
[9] Campbell, W. N. and Black, A.W. "CHATR a multi-lingual speech re-sequencing synthesis system". Technical Report of IEICE SP96-7, 45-52, 1996.
[10] Campbell, W. N., "Speech & expression; the value of a longitudinal corpus", pp.183-186 in Proc Language Resources and Evaluation Conference (LREC-04), Lisbon, Portugal, 2004