User Interface for an Expressive Speech Synthesiser

Nick Campbell
ATR Human Information Science Laboratories,
Keihanna Science City, Kyoto, Japan
nick@atr.jp

1 Introduction

The term 'allophone' is used in phonetic research to describe the variations in acoustic characteristics of a given phone in different contexts. A typical example is the dark and light variants of labials and nasals in onset and coda positions. The initial and final consonants in words such as "mum" and "lil" are considered to be the same in spite of great differences in their physical characteristics [1]. This paper proposes the term 'allophrase' for words or phrases that are considered to be the same in spite of differing acoustic characteristics, and examines the criteria by which an appropriate token may be retrieved from a database for use in concatenative speech synthesis. Whereas the allophone can be succinctly described by its phonetic context, the allophrase is more dependent on discourse context and interpersonal factors. Unlike the allophone, substitution of a different allophrase can result in the perception of a different meaning for an utterance. Just as an allophone is not distinguished except by its phonetic context, so these allophrases are usually transcribed identically, yet carry different meanings depending on their acoustic realisations.

2 Expressive conversational speech

In conversational speech, both the listener and the speaker strive to maintain social relationships at the same time as exchanging propositional content [2]. The voice and speech prosody are controlled to signal not just grammatical and semantic relationships, but also discourse and interpersonal factors [3].

Non-verbal utterances are common and they are used for signalling the paralinguistic information. For example, the word 'yes' (typically 'yeah' or 'yup' in friendly conversational situations) can function as a back-channel for showing affective states such as agreement, understanding, hesitation, doubt, sarcasm, participation, etc., in addition to its standard lexical meaning in propositional utterances. These two forms of usage can be distinguished as either of I-type (information) or A-type (affect) utterances [3]. Allophrases are usually A-type utterances, though some may also have I-type versions.

Whereas I-type utterances can be sufficiently described by their text transcription alone, A-type utterances also require both prosodic and voice-quality information before they can be successfully interpreted by the listener. Similarly, for synthesis, a text input may be adequate for propositional content, but markup is required for expressing affect. For conversational speech synthesis, whether it is for use in customer-care applications, speech translation, or support for the speaking-impaired, fine control of not just linguistic but also paralinguistic information will be required. Current markup conventions (such as SSML [4]) are too low-level to be of use for allophrase discrimination so we describe below a framework for the specification of A-type segments in conversational speech.

表現豊かな音声合成入力インタフェースの提案
○　ニック　キャンベル
ATR-人間情報科学研究所、けいはんな学研都市、京都 .

3 Synthesising conversational speech

Previous work [5,6] has described a novel method for concatenative synthesis using a large unlabelled speech corpus. Recent extensions to that work have resulted in methods for detecting repetitive segments of the speech [7] and for training tree-based models for labelling their affective characteristics [8].

This section addresses the issue of specifying input to the synthesiser so that such segments may be retrieved for concatenation. The 'tap2talk' interface [9] was proposed for the synthesis of emotional speech, and is available at [10] (for illustration only) with an i-mode implementation of the interface. We describe here an extension to that input device and a framework for categorising speech utterances accordingly.

1 utterance specification
2 target waveform synthesis
3 selection from corpus
4 filtering by features
5 output waveform generation

Figure 1: Processing flow for conversational speech synthesis. An intermediate waveform (2) is used to retrieve candidate segments and a post-filtering (4) is applied to retain only those candidates having the appropriate affective characteristics.

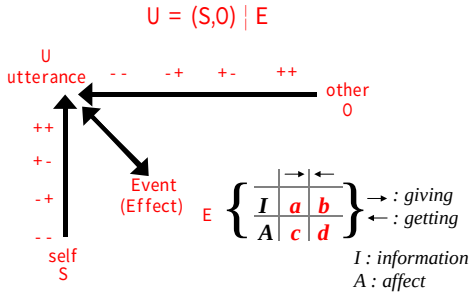

Figure 2. The 3-dimensional SOE framework for determining the realisation of an utterance.

Althouth the utterance specification for this type of synthesis does not need to be as precise as that for text-to-speech, it is instead necessary to specify three higher-level contextual characteristics, as in figure 2 [3], where *U* represents an Utterance, *S* and *O* are self, and other, respectively, and *E* represents an event (or speech act) in the case of speech production, or an effect in the case of perception. Self (4-levels in this implementation) represents 2 features found to be dominant in speaking-style control: mood and interest. If motivation or interest in the content of the utterance is high, then the speech is typically more expressive. If the speaker is in a good mood then more so. If the listener (other) is a friend, then the speech is more relaxed, and in a friendly situation, then even more so. The utterance is realised as an event (E) taking place within the framework of mood and content (Self) and friend and friendly (Other) constraints implemented here with 4 levels of activation each.

4 Utterance as event

The amount of choice in generating an utterance is usually very limited in text-to-speech synthesis, and highly constrained in concept-to-speech. However, for conversational speech synthesis, an utterance can be defined (as detailed above) as the result of several higher-level factors. A human speaker has a very wide choice of alternatives for social or A-type utterances. For example, “Hi”, “Hello”, “Hey”, “Good morning”, “How are you”, “How do you do”, “Ah”, “Oh”, “Hi there”, even “Nice to see you the other day”, are all simple greetings – the speech act is the same, but the style and expressiveness vary – not within a single dimension (e.g., that of politeness) but within the framework described in figure 2. The four classes of Event constrain the brackets of alternatives, and the SO settings narrow down the choice within them.

For computer speech synthesis using a telephone keypad instead of a keyboard, the factor combinations can be selected using three buttons for the SOE factor, three for the SVO (subject-verb-object) factor described in [9], and three for extralinguistic features such as language (L), speaker personality (V), and energy (E) as in figure 3, which extends the tap2talk interface [9] shown in figure 4.

5 Event-based synthesis

To map from the selector settings to the speech waves, we use generalisations of the acoustic features from models trained on data gained from human perception experiments that required native listeners to identify the affective traits [8]. Given that for highly frequent A-type utterances we can use the entire phrase as a segment for synthesis, there is no concatenation involved. Instead, we need to identify the segment having the most appropriate speech characteristics from amongst a very large number of similar candidate segments. Since the data is not labelled, we rely on generalisations of the acoustic features such as speaking rate, prosody, voice-quality, etc.

Once the utterance is specified by choice of icons in the SVO section, the SOE section, and the LVE section, the nature of the speech is highly constrained. The closest allophrase matching the specification is then retrieved from the database and replayed intact. The skill in selection is to find the one from many clustering in the same space which best matches the intended utterance characteristics. This may require selecting one having different text but similar intended meaning, which reflects the choices human speakers also make.

6 Conclusion

Earlier work with the CHATR [11,12] system facilitated extralinguistic control for speech synthesis. The present work extends this model to include paralinguistic controls based on the study of a large corpus of spontaneous conversational speech. The present paper describes a framework for reducing the high dimensionality of this space into a small number of alternatives so that the process of determining both an utterance and its realisation style can be carried out using a small numerical keypad. The paper lacks an evaluation section, partly because this work is still in progress, but also (since the A-type utterances are replayed intact from the corpus and their naturalness can be guaranteed) objectively judging their appropriacy is currently beyond our technical ability. This is being carried out as current and future work.

Figure 3. The 3-dimensional SOE framework for determining the realisation of an utterance. The SOE row is added to the SVO and LVE rows described in [x]. The i-appli is under development.

Figure 4. The prototype i-appli for tap2talk and natr which can be found at http:feast.his.atr.co.jp/i is currently limited to DoCoMo 503i handsets only.

Acknowledgements

The author is grateful to the JST-CREST and members of the ESP project. This work is also partly supported by the Telecommunications Advancement Organisation of Japan.

References

[1] Martinet, A., “Function, structure, and sound change”, Word 8.2:1-32, 1952.
[2] Campbell & Mokhtarri, “Voice quality, the 4th prosodic dimension”, Proc ICPhS 2003, Barcelona.
[3] Campbell, “Expressive speech – simultaneous indication of information and affect”, in Feschrift for Wu Zogn-ji, Beijing 2004.
[4] SSML: www.w3.org/TR/speech-synthesis
[5] Campbell & Mokhtari, “Using a non-spontaneous speech synthesiser as a driver for a spontaneous speech synthesiser”, IEEE w/s on Spontaneous Speech Processing, Tokyo, 2003.
[6] Masaki, Kashioka, Campbell, this volume.
[7] Ashimura, Campbell, this volume.
[8] Campbell, & Erickson “What listeners hear; the interpretation of affect in conversational speech”, Journal of the Phonetic Society of Japan, April 2004 (forthcoming).
[9] Campbell, “tap2talk: an interactive interface for large speech corpora, Proc ASJ Spring mtg, 2003.
[10] The ESP web pages: http:feast.his.atr.co.jp/
[11] Campbell, “Synthesis Units for Natural English Speech” 電子情報通信学会技術研究報告 SP 91-129 1992.
[12] Campbell, “CHATR: A High-Definition Speech Re-Sequencing System”, in Proc ASA/ASJ Joint meeting (Hawaii) 1996.