MULTI-LINGUAL CONCATENATIVE SPEECH SYNTHESIS

Nick Campbell
ATR Interpreting Telecommunications Research Labs.
2-2 Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-02, JAPAN
nick@itl.atr.co.jp, www.itl.atr.co.jp/chatr

ABSTRACT

This paper describes a method of concatenative speech synthesis that makes use of 3-dimensional labelling of speech, and shows how this can be applied to the synthesis of both mono-lingual and foreign-language speech. The dimensions encode phonetic, prosodic, and voice-quality information in order to fully describe the acoustic characteristics of each speech segment.

1 INTRODUCTION

CHATR speech synthesis produces speech without signal processing. It re-uses segments of natural speech that have been selected from a large corpus by consideration of their prosodic and acoustic characteristics and simply concatenates them.

CHATR represents not just the 'removal of signal processing' in favour of prosody, but rather a new description of the relevant information in speech that enables synthesis with natural voice quality.

Previous assumptions about the nature of speech, that formed the foundation of many speech synthesis applications, were largely based on Fant's innovative work [4] using one-dimensional source-and-filter models which implicitly reinforced the phonemic view of speech and, in turn, relegated prosody to a secondary or supra-segmental position from which it could then be modelled as an independent process.

Experiments with CHATR have confirmed that an 'integrated' view of speech, coded as a three-dimensional information system, allows for a more efficient representation of the information contained in the speech signal and leads in turn to more natural-sounding speech synthesis.

2 3-DIMENSIONAL SPEECH

The first dimension of speech information must of course be phonetic; it describes the nature of the segment, such as is governed by its place and manner of phonation. The inherent acoustic characteristics of each speech sound thus defined are further moderated by context, with the strongest effects coming from immediately neighbouring segments, although with influences also being observed from more distant segments, depending in turn on the nature of their own articulation and contexts.

The second dimension of information in the speech signal is prosodic. The acoustic characteristics of the segments change significantly according to their position in the utterance and according to the stress and intonation patterns that arise from differences in the meaning ascribed to their words. Focus, stress, emphasis, reduction, pre- or post-boundary contexts etc., can affect the acoustic characteristics of the speech segments almost as much as differences in their phonemic type and context.

The third dimension with which to categorise speech is phonation. The same sequence of sounds with the same prosody can carry different pragmatic or functional interpretations according to the manner of phonation, or 'tone-of-voice', under which they are produced. For example, breathiness is typically associated with a softer style of speech, and harshness with aggression. Far from being simply personal or idiosyncratic characteristics, these different phonation styles can be usefully manipulated for effect in human interaction.

3 HIGH-LEVEL LABELLING

Phonetic knowledge is already well developed and can be considered as a mature branch of speech science, linguistics, or engineering. Prosodic knowledge on the other hand, while also thoroughly investigated, is still in a development stage, and there are currently several controversial or competing approaches to account for prosodic variability. The signal characteristics and social uses of voice phonation however, still remain largely unexplored outside the domain of speech science.

3.1 Reducing Complexity

Just as the complexity of the acoustic space can be reduced by use of phonemic labels, so the complexity of the prosodic space can be reduced by use of higher-level labels that characterise the main features. There is a growing consensus (as evidenced by e.g., the ToBI movement [6]) that we can consider the primary dimensions of the prosodic space as being defined by stress, tone, and boundary.

For an initial approximation, these dimensions can be considered as having binary status with a neutral, or unmarked state indicating a 'don't-care' category in which the prosodic characteristics can be inherited from neighbouring contexts. We can refer to this two-state-plus-neutral space as '*binary+*' and can visualise it as a sigmoid function having a wide unstable area. This simplifies discussion about whether the prosodic features are really binary (or should rather be considered as scalar) by reducing it to one of 'degree of slope' for the transition.

The inclusion of the unmarked state is crucial to specifying acoustic characteristics, however, as it allows us to consider the status of neighbouring elements as part of the definition of any given segment. If a segment is marked with a feature, we can consider its influence to spread over neighbouring segments if they are themselves unmarked. The degree of this spreading is a matter for continuing research.

3.2 A Syllable-level View

Because prosody is primarily signalled by the sonorant part of the speech, we can consider the syllable to be the smallest unit that is marked for prosody (not to be confused with 'subject to the effects of prosody'). We can thus simplify the notation of prosodic effects by categorising each syllable according to its status with respect to *stress* or prominence (plus, minus, unmarked), *tone* (high, low, or unmarked), and *boundary* (pre-, post-, or unmarked). Thus the wide range of possible prosodic variations can be reduced to a simple three-dimensional combination, having 27 possible states, per syllable.

By labelling each syllable as having one of these states, and considering each as contextually influenced by a window of similarly marked neighbours, we can capture the significant prosodic effects on its acoustic features. Assigning weights according to the prosodic characteristics when selecting segments for concatenation in CHATR allows us to test this integrated view of speech information. Results confirm that enough of the information is captured by the combination of prosodic and phonemic features to enable us to reduce signal processing to a minimum or even to eliminate it completely [1].

3.3 Vocoid and Contoid Sequences

CHATR labels speech using the phonemic symbols produced by a pronunciation dictionary, and concatenates phone-sized waveform segments to generate novel utterances. In order to find the segment sequences which are as close as possible to the intended target, to replicate human speech, we need to note only the perceptually relevant acoustic variation in each.

Pursuing the syllable-based view of speech, we can further reduce the number of phonetic variants by considering each syllable to be defined primarily by its sonorant vocalic nucleus, and coloured, secondarily, by the contoid nature of its onset and offset characteristics. Since the vocoid space can be well characterised in two primary dimensions (front-back and open-close, each with an unmarked neutral area or schwa) then we can also make use of a *binary+* definition of the acoustic variation in the syllable nuclei. Using a further feature to describe the contoid tier, and one more for phonation type, completes our index into the significant variation underlying, or required to define, a given portion of speech.

The vocoid 'carrier' is of course also influenced by the contoid perturbations, but they are more fixed in their characteristics, and serve rather to specify the 'lexical' meaning of an utterance, whereas the vocoid tier carries more of the para-linguistic information which signals its intended interpretation.

The different consonants have different degrees and direction of interaction with their neighbouring sounds, and this effect too can be considered as *binary+*. For example, plosives strongly affect the formant structure of the neighbouring vowels according to their place of articulation but the fricative /h/, on the other hand, is usually more affected by the formant structure of the vowels themselves. Strength of contoid influence can thus be (positive, negative, or unmarked).

This model does not deny the contribution of contoid sounds to the speech signal but rather delegates them to a secondary place in terms of effect, for by giving priority of definition to the vocoid sequence we can better capture the varying nature of the underlying signal. And by use of high-level features, we capture the lower-level acoustic effects implicitly.

4 EXPRESSING EMOTION

By keeping speaking style relatively constant across the entire source-speech corpus, little use has previously been made of phonation-style or voice-quality information in CHATR experiments. Variation in this dimension has been a source of 'noise' in the concatenative synthesis. In this section we show how phonation can be of great importance in the signalling of 'affect' in speech.

As an example, we will consider the case of emotional speech. It is sometimes questioned whether there is a need for synthetic speech to show emotion, but when a synthesiser is used as a prosthetic device [5] or in a translation system [7] then we believe there is no alternative.

There are clear prosodic correlates of emotion, such as reduced pitch-range or slower speaking rate in sad speech, but simply reproducing the intonational and durational peculiarities is not sufficient to replicate the signals that are usually present in human emotional speech. The quality of the voice changes as much as the prosody under conditions of marked emotion, but rather than attempt to model this by parametric methods, such as would be needed for direct modifications to the spectrum, we can instead label their primary characteritics as a feature on the syllable and then select only those segments having the desired characteristics. This, of course, requires a large source corpus.

In order to test labelling of phonation style, we created three separate corpora, consisting of about an hour of speech each, which were highly marked for either joy, anger, or sadness. By switching between these we were able to convey the intended emotional attributes even in semantically neutral utterances [5].

It remains as future work to distinguish automatically between the acoustic characteristics of the different emotions, but if this can be done (presumably using parametric measures) then we can merge the three corpora into one, and select appropriate emotionally-coloured speech segments using context-specific features similar to those that we currently use for distinguishing prosodic differences.

5 MULTI-LINGUAL SPEECH

In a closing panel session of the 1996 ICSLP, Sadaoki Furui made the challenging observation that computer processing of speech should not just be required to simulate human performance, but that the machines should be expected to offer something above and beyond the level that human performance is capable of.

With this goal in mind, we have been exploring applications of CHATR that extend human performance. The first such application concerns multi-linguality. Many speech synthesisers have been capable of multi-lingual output but this has usually been in a mechanical-sounding voice that is 'owner-less'. Since CHATR produces speech in the recognisable voice of a known person, it offers the potential to extend that person's apparent abilities into the realm of multi-linguality. By offering this ability to the voice of a young child, we are perhaps meeting Furui's expectations [3]. [SOUND 0024.01.WAV][SOUND 0024.02.WAV]

5.1 Foreign-language Synthesis

Language is only partly dependent on the speaker in the CHATR system. For example, when a Japanese person speaks in English, unless they are particularly fluent, the range of variability in the resulting vowel space will probably be closer to that of the mother-tongue than to that of a native-speaker of English. This is one of the causes of 'foreign accent'; the restricted range of prosodic variation, in accordance with mother-tongue patterns, is another.

However, many people can successfully communicate in foreign languages without really departing far from the prosodic and phonemic spaces of their native languages. They do this by re-sequencing their own familiar speech sounds in an order appropriate to (at least) the lexis of the target language.

By mapping from the phone sequence predicted for synthesis in one language to the phone-set used to label the speech of another, we can produce foreign-language speech using the voice of any speaker. In these examples we use the voice of a small Japanese child to speak in English ([SOUND 0024.03.WAV][SOUND 0024.04.WAV] greeting) and Korean ([SOUND 0024.05.WAV] [SOUND 0024.06.WAV] explaining the technical processing within CHATR).

5.2 Two-stage Language Mapping

At ATR, we are researching speech synthesis for use in translated speech. It is beneficial to produce output speech using the voice of the input speaker, but not to give an impression of language incompetence. To reduce the 'accent', we adopt the following two-stage process: ([IMAGE 0024.01.GIF] schematic).

We first synthesise the target speech using the voice of a native speaker of the target language and then, using the resulting acoustic waveform (or its cepstral representation) as a physical target, select speech segments from the pre-stored voice database of the input speaker by minimising a physical distance measure.

This use of a physical target for unit selection is not feasible in monolingual synthesis since, by definition, if the utterance existed in a suitable form there would be no need to synthesise it. However, by making use of a native-speaker's speech as an intermediate target, we can narrow down the selection of speech segments to match the spectral characteristics of the native, thereby making use of the natural variation in production that could not be accessed through label information alone.

6 DISCUSSION

There has been much discussion about the distinctions between 'measurers' and 'modellers' in the world of prosody. We believe that CHATR offers a bridge between these two supposedly opposite approaches, since by the quality of the synthesis we can judge the adequacy of the underlying model, which in turn is derived from (or in the case of re-sequencing, 'dependent on') the measurements made on the original corpus.

By simply labelling a speech corpus in terms of its component segments, with phonetic labels, we are able to reproduce only the 'text' of an utterance. With the addition of prosodic labels we can then indicate its intended 'meaning'. By the further addition of phonation labelling, we can reproduce its 'nuances'.

6.1 Phonology of Acoustics

The granularity of the labels determines the generality of the model. Without inclusion of higher-level information, the phonetic labels would have to be very 'narrow' to be able to specify the minor but significant changes in the acoustics that accompany e.g., stress, or pre-boundary position. However, by incorporating such higher-level influences as a further dimension of labelling, we are able to select units from the database that have appropriate acoustic characteristics to signal the required prosodic event even when using 'broad-class'

canonical phone labels. By acknowledging the dependency on prosodic as well as on phonetic context, we are able to capture the finer distinctions with simpler labels.

The inclusion of phonation style is necessary when processing larger speech corpora. What remains now is to be able to label (i.e., detect and discriminate) such information automatically. For the first two dimensions, this is already known technology. The interesting challenge that lies ahead is to recognise voice-quality differences without having to listen to the speech.

Clearly, there is also a case for larger prosodic domains than the syllable, as evidenced by annotation at the intonation-phrase and utterance levels in ToBI (marked by '*'. '–', and '%' respectively), but until very large single-speaker speech corpora become available, we will not have the materials needed to test such an implementation.

6.2 Interactive CHATR

Although not widely advertised, CHATR has been available in interactive mode on the world-wide-web since April this year [2], but we have strong reservations about such access to the technology. On the one hand, it is essential that potential users be able to test its output for themselves, but since we make use of recognisable people's voices (ambiguity intended) then we need to take particular care about making them freely available. Future work should perhaps include watermarking of generated speech, but for the present we rely on software logs for accountability.

6.3 Evaluation of Speech

A recent announcement from a well-known TTS Workshop Evaluation Committee stated: " ... in the typical case the many voices all ride on top of exactly the same software, and hence *are not really different after all*" (my italics). We now have more than a hundred speech corpora processed for CHATR synthesis, and no two voices are the same. Judging them is like judging people; we can favour one voice over another but we cannot say that A has a better voice than B. If future synthesis evaluations are to grade voices, perhaps they should start by evaluating the extent to which the voices can portray the depth of meaning that the human voice is capable of. This would be a test not just of the synthesis algorithms, but also of the adequacy of database labelling.

The Turing test must ultimately be the best form of evaluation for speech synthesis. If a human listener believes that another human is speaking, then the system can be said to have passed this test. Many present synthesis systems might pass such a test if the amount or type of speech could be constrained, but probably none would be able to exceed even a minute of free conversation. So the more interesting question for the current technology is what limitations we can reasonably put on the Turing test to make it a useful measure of synthesis quality.

7 CONCLUSION

In this paper we presented a framework for the labelling of speech information that allows richer representation of the underlying information, and we made the claim (confirmed by CHATR synthesis) that this 3-dimensional representation of speech information is adequate for the modelling of human speech in a wide range of situations.

By this higher-level indexing of the speech corpus we were able to eliminate signal processing from the synthesis process and were thus able to reproduce the original speaker's voice with a high degree of fidelity. This method extends traditional synthesis capabilities to enable reproduction of children's voices in the just same way as those of adults.

We further showed that the voice of a known speaker can be re-used not just for monolingual synthesis, but also to reproduce sounds in foreign languages, and we described a two-stage algorithm that provides finer control of accent and pronunciation than previous phone-based descriptions.

Finally, we showed that emotion can be signalled not just by intonation, as has previously been held, but by 'tone-of-voice', justifying the inclusion of the third dimension of speech information.

The CHATR method of voice reproduction relies on the availability of large well-balanced corpora of speech data for a given voice and speaking style. It does not have the flexibility of conventional parametric synthesis, but for a closed-domain task it has proved capable of extremely high-quality speech and has passed the Turing test for synthesis in at least one application[1].

References

[1] http://www.itl.atr.co.jp/chatr.

[2] http://www.itl.atr.co.jp/chatr/interactive.

[3] http://www.itl.atr.co.jp/chatr/j_tour/yuto2.html.

[4] Fant, G. (1991) "What can basic research contribute to speech synthesis?", J. Phon. 19, 75-90.

[5] Akemi Iida, "The acoustic nature and perceptual impression of a corpus of emotional speech", Proc. of ICSLP98, (December 1998).

[6] Silverman et al. (1992) "ToBI: A standard for labeling English prosody", Proc. ICSLP'92, Banff, 867-870.

[7] Toshiyuki Takezawa, Tsuyoshi Morimoto, Yoshinori Sagisaka, Nick Campbell, Hitoshi Iida,Fumiaki Sugaya, Akio Yokoo, Seiichi Yamamoto: "A Japanese-to-English Speech Translation System: ATR-MATRIX," Proc. ICSLP98, (December 1998).

[1] Omron has already incorporated CHATR technology in a prototytpe telephone-shopping system where synthesis is combined with recorded utterances for the generation of open-class items such as credit-card numbers