Design of Speech Corpora for use in Concatenative Synthesis Systems

Nick Campbell
ATR Interpreting Telecommunications Research Laboratories
Hikari-dai 2-2, Kyoto 619-02, Japan.
nick@itl.atr.co.jp, www.itl.atr.co.jp/chatr

Abstract

This paper describes some speech corpora that are currently being produced at ATR in Kyoto, Japan. It also addresses the issue of maximising the usefulness of the speech corpora that are currently being planned and created in various countries around the world. Specifically, it argues for taking into account the needs of the widest possible range of users when designing and recording such corpora.

1 Introduction

Most of the large speech corpora that are currently being distributed were designed for use in the training of speech recognisers, and include data from a very wide variety of voices and different recording conditions for the development of robust speaker-independent recognition systems. Individual samples from a single speaker are typically limited in duration to the order of a few minutes each.

However, the needs of the discourse, synthesis and prosodic analysis communities are for similar but longer samples of speech, to enable an analysis of features such as variation in speaking style, turn-taking, and paragraph-level prosodic characteristics. I argue here that consideration of such needs would place only small demands on the design and collection of future speech corpora, but that the small changes in data categorisation could greatly benefit the wider communities.

2 Corpus-based Synthesis

By directly concatenating raw waveform segments from the speech corpora, without recourse to signal processing for modification of their prosody, realistic-sounding machine-generated speech can be created. However, the cost of such synthesis is that the source corpus must be big enough to contain many examples of all the basic speech sounds in each typical prosodic context from every candidate speaker.

At ATR we have been developing advanced corpus-based speech synthesis techniques, and have recorded and labelled many single-speaker corpora.

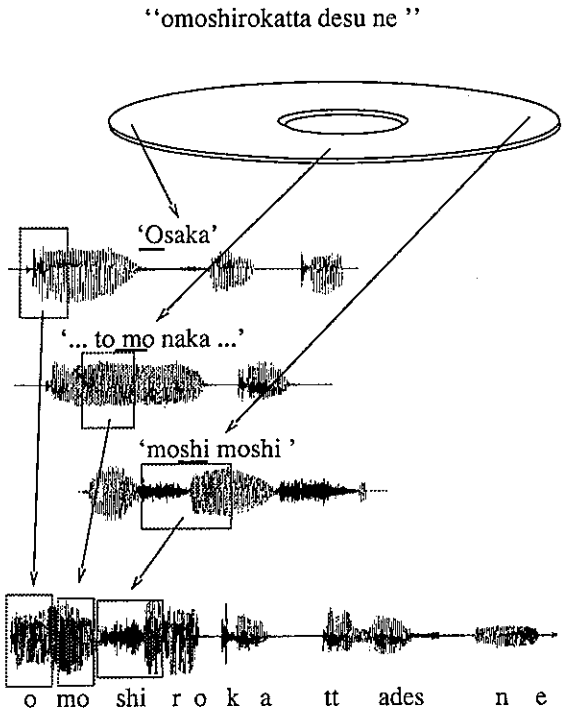

Figure 1: Selecting speech waveform segments from different places in a corpus to create novel utterances using the voice of the original speaker.

We base our techniques on the re-use of large speech corpora, typically requiring at least an hour of speech from each voice for high-definition 'personality-preserving' speech[1]. This section presents a brief summary of our synthesis work and a description of the types of corpus we have found useful.

To date we have collected 97 large speech corpora for synthesis, from the voices of 59 different speakers, in six languages. The corpora vary in style from readings of lists of isolated words, through readings of phonemically-balanced sentences, short stories and web pages, to free spontaneous monologues and conversation. Durations of recordings from a single speaker typically vary between twenty minutes and four hours, though we are currently analysing an eight-hour and a sixteen-hour corpus from one speaker.

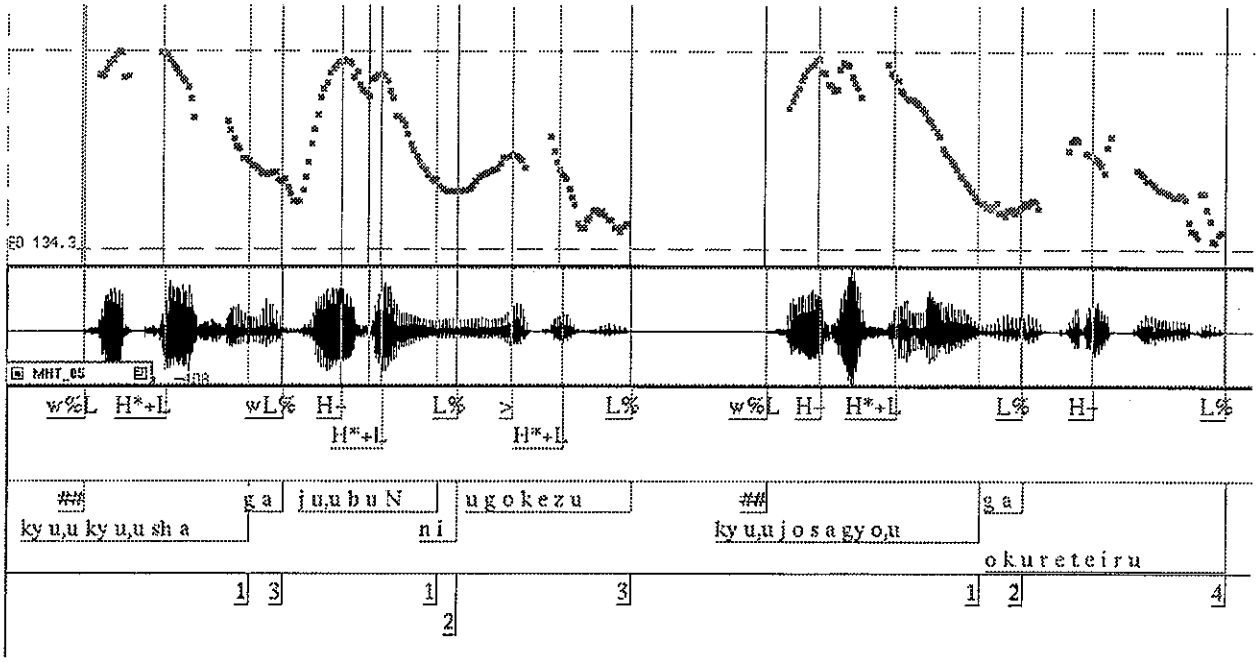

Figure 2: Labelling a speech corpus for phonemic and prosodic information about each segment.

We originally used lists of the 5,000 most common words to collect speech samples representative of the sound combinations of a given language, but the intonation produced when reading such isolated-word lists is not at all representative of the intonation required for continuous speech and the resulting speech sounds themselves tend to be over-articulated, presumably to emphasise the contrasts between the words, in the absence of a defining textual or semantic context.

In an intermediate stage of CHATR development, we tested speech from readings of sets of 500 phonemically-balanced sentences for use as source units for the concatenative synthesis. However, as these texts contained many items that were difficult for most of our speakers to pronounce (having been inserted to ensure full coverage of all triphone combinations), we found that the stress and tension (or lack of interest) during the recording session remained present in the voice and resulted in unnatural 'flat-sounding' concatenated speech.

We currently ask our speakers to bring a novel or short-story of their own choice when recording a new voice. This leaves the matter of phoneme balance almost to chance but, surprisingly, analysis of the resulting corpora has shown little diffference between the phonemic distributions of long randomly-chosen text and those of more carefully designed corpora, once a certain size threshold has been achieved.

The relaxed state and 'interested' tone-of-voice that arises from reading of continuous text produces a pleasant and useful voice quality in the corpus.

2.1 Random-access segment replay

Our method of indexing speech segments according to their joint phonological and prosodic attributes allows us to select candidate waveform segments with sufficient accuracy for concatenative synthesis without signal modification (see Figure 1). By specifying the variability in these two dimensions we are able to characterise the speech of any given speaker, but only for one given mode of speaking.

If the speech corpus has been collected over a period of several days or weeks, or includes several different types of speaking style, then the likelihood of different phonation styles or emotional attitudes increases, with consequent discontinuities in the voice-quality of the output synthesised speech, so the need for a third dimension of indexing arises. Research into such changes in voice characteristics within data from a single speaker is currently under way.

When categorising the vocal variation in such richer corpora, we need a three-dimensional integrated index which combines phonetic, prosodic, and phonatory classes. Using the above features we are testing ways to select appropriate segments for concatenation to produce speech that not only reproduces the intended focus and prosodic bracketing of a spoken utterance but also takes advantage of voice quality to show emotion.

Examples of concatenative waveform synthesis showing three different emotional states (sadness, anger, and joy) using the voice of a single speaker can be found at [3]. It can be seen from these that the emotion remains in the speech even for the production of a neutral utterance.

2.2 Labelling a speech database

An example of the phonological and prosodic labelling is shown in Figure 2. The minimal requirement for automatic labelling of speech data is an orthographic rendering of the text of each utterance, from which phoneme sequences can be determined and aligned using a combination of synthesis prediction and hidden Markov techniques. If only speech data alone is available, then we estimate a week of human labelling time per half-hour of speech. Once a phonemic index into the speech data is available, then the prosodic characterisitics of each phone-sized segment of the speech can be determined and a full index into the corpus is prepared.

To synthesise a novel utterance, we pre-select a limited number of candidate units from this full index and find the best sequence for concatenation by Viterbi alignment, according to the two criteria illustrated in Figure 3. The optimal sequence of speech waveform segments must closely match the required prosodic targets for the utterance while at the same time fitting smoothly together so that any discontinuities at their join points will be imperceptible.

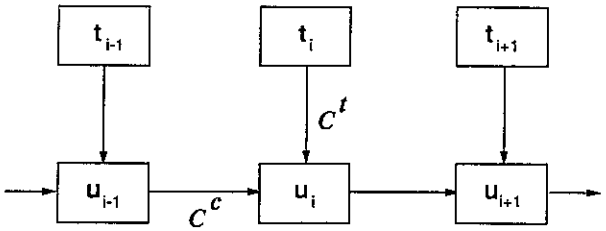

Figure 3: Two functions to select the speech segment that is closest to the target prosody while concatenating smoothly with its neighbours.

2.3 Three aspects of an utterance

An utterance has three primary defining characteristics: a speaker, a language, and an intended meaning (Figure 4). While differences in meaning may be the hardest to analyse and control for speech synthesis, we can freely interchange the other two dimensions by corpus switching.

For example, two people saying 'Hello' are probably using the same language with the same intended meaning, but if instead one were to say 'Bonjour', then only the language dimension would have changed. Similarly, a speaker saying 'hello' on two different occasions may be performing two different functions (e.g., greeting vs. exclamation) and thus changing the 'meaning' dimension.

For research applications in interpreted telecommunications we are particularly interested in using the voice of one speaker to produce utterances in the language of another. For this, we need access to large single speaker corpora from many languages.

3 Multilingual Corpora

Selection of a speech segment sequence for synthesis in CHATR is performed by comparison of the features of each candidate unit against a vector of higher-level or abstract features specifying the desired characteristics of the utterance to be synthesised. However, in certain cases it is possible to use very low-level or physical acoustic characteristics as targets specifying the candidate unit.

For example, if we have a sequence of cepstral vectors specifying how a native speaker of a language would produce a given utterance, then we can use this sequence as a direct target for the selection of speech waveform segments from the voice of a non-native speaker, in order to most closely represent the utterance as produced by a native. In this way, we can both protect the identity of the original database speaker and produce high-quality multi-lingual synthesis for speech translation. For further details of this technique, please refer to [4, 5].

To date we have processed (with permission) two non-ATR commercially-distributed CD-Rom speech corpora, one of German and one of Chinese. Each presented language-specific problems associated with phonemic labelling, but the experience has confirmed our views of the universality of speech and corpus processing.

We use dictionary-generated and forced-aligned phonemic labelling of the speech to produce an index of prosodic characxteristics per phone-sized speech segment. The distributed copora were labelled with different degrees of accuracy and at different levels of abstraction, so both required manual adaptation to prepare a set of labels before initial processing could begin, but both were successfully handled by CHATR and the resulting voices are illustrated on our web pages.

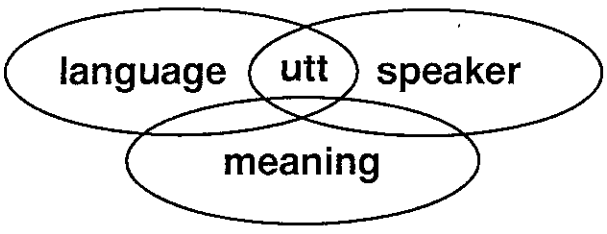

Figure 4: Three aspects of a spoken utterance. Each can be manipulated independently of the others

4 Design of Speech Corpora

Evaluation of speech corpora must of course be primarily dependent on the applications for which they were designed, but because of the difficulty of preparing and annotating such resources, and of making them widely available, we would encourage inclusion of some extra samples outside the original design specifications in order to widen their potential for use.

Neither of the non-ATR corpora which we tested were designed for large-corpus synthesis, and in both cases there were only two speakers (one of each sex) who presented enough samples for our processing, but we are very grateful to the designers for including these speakers.

Our research interests include dialogue prosody, speaking style, and voice characteristics of extended discourse, but unlike those in the speech recognition community who prefer wide coverage of speaker types, we can be content with only a small number of sample speakers.

Our criteria for evaluation of available speech corpora start with scope of segmental coverage but also include prosodic variation, specified in the two dimensions of salience and segmentation, requiring each vowel in positive, neutral, and negative salience, and in pre-boundary, post-boundary, and boundary-neutral contexts.

In addition, we are also interested in understanding the dimensions in which voices vary over time. For this, we require speech collected at different occasions and speech produced under different emotions. Clearly, this has to be speech from the same speaker to enable meaningful analysis.

In summary, we encourage the collection of larger amounts of speeech data from a limited number of speakers, preferably covering all age ranges and both sexes, across a range of different languages. The effort of recording and producing CD-Roms is negligible compared to the effort of segmentation and labelling the speech data. This latter task can be performed collaboratively across the world, perhaps by requiring the sharing of derived information as a condition for use of the speech corpora.

5 Copyright issues

I have been advised (and would be pleased to hear of any contrary opinions) that there can be no copyright on the sounds of a speaker's voice. Legal parallels are drawn with other component elements such as colours in a painting or words in a book. Since it is only the original combination of these basic elements that can be subject to legal protection, the basic elements themselves are considered to be in the public domain.

We should work for changes in this legal situation if raw waveform synthesis methods like CHATR are to become more widespread, but for the immediate future, care must be taken that corpus speaker rights are not abused. Novel combinations of the individual sounds in a speech corpus may be legally equivalent to original works of art but if they were to be mistaken for original speech spoken by a known speaker then they might cause personal offense or embarrassment.

By mapping from speech generated using the voice of a corpus speaker onto the voice of a CHATR-registered speaker we can preserve the corpus-generated synthesis as an internal and intermediate element of the final synthesis and thereby avoid any potential for infringement of ethical or legal rights.

6 Conclusion

This paper has presented examples of the multilingual application and re-use of existing corpora for synthesis both in the original language and across languages. Examples illustrating the potential of large-corpus-based speech synthesis can be found at www.itl.atr.co.jp/chatr,

It would be of great interest to develop the system further, using speech data from other languages and from a wider variety of speaking styles, but there are currently very few corpora available which contain enough speech data from a single speaker to be of practical use. Perhaps when future data collections are planned, there can be allocated at least one person of each sex who will speak for sufficient time to provide enough prosodic and phonemic variation for this and other similar research to be carried out.

In order to protect the rights of the original speakers, who may not have been informed of such applications of their speech data, we are exploring techniques to map from the speech of the various native speakers and languages on to that of a known and registered voice for use in research towards practical applications of our automatic speech translation algorithm.

References

[1] CHATR **Speech Synthesis:** http://www.itl.atr.co.jp/chatr ATR Interpreting Telecommunications Research Laboratories, Kyoto, 1997.

[2] **The** CHATR **User Guide:** http://www.itl.atr.co.jp/chatr/manual

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

[4] Campbell, "Large-scale Single-speaker Speech Corpora", Proc Oriental COCOSDA, Tsukuba, 19989.

[5] Campbell, "Multilingual synthesis", Proc ICSLP-98 (submitted).