Large-scale Single-speaker Speech Corpora

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 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.

For example, most of the large speech corpora currently being distributed were designed for use in the training of speech recognisers, and include data from a very wide variety of speakers for the development of speaker-independent recognition systems. Individual samples from any 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. As an illustration of this wider applicability, we present samples from a speech synthesis system that makes use of large single-speaker corpora for source units, allowing multi-speaker synthesis across languages.

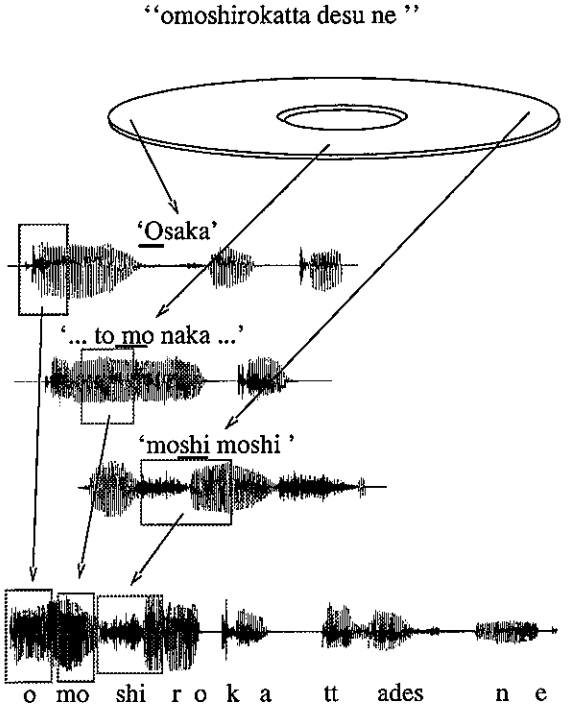

Figure 1: Selecting speech waveform segments from different places in a corpus to create novel utterances using the voice of the original speaker.

1 Introduction

Many large corpora of speech are now being collected and distributed, but the design of these is often dictated by the needs of speech and speaker recognition technologies and as a consequence, although they typically contain several hours of speech, there are rarely more than a few minutes of speech samples from any single speaker. Since the collection and annotation of large speech corpora is an expensive and time-consuming process, we should take extra consideration when designing and evaluating such materials so that the needs of the widest possible range of scientific and industrial communities can be met.

As the demands of speech analysis change, we should consider collecting materials that not only meet present requirements but that will serve future needs as well. By doing so, such disciplines as dialogue research, speech synthesis, and prosodic information processing can also benefit from the corpora being currently prepared. In many cases, only minor changes would be necessary, for example using longer reading passages for one or two speakers and including more conversational and spontaneous speech materials, but the means for making such requirements widely known are not yet well enough established.

COCOSDA was established to encourage and promote 'international interaction and cooperation in the foundation areas of spoken language processing', and it emphasises the importance of collaboration which transcends national boundaries. COCOSDA is therefore the ideal forum at which such requirements should be discussed. However, as COCOSDA is currently organised, the three working groups (Recognition, Synthesis, and Labelling) offer little interaction between the different communities, which are perceived as having non-overlapping needs.

2 Corpus-based Synthesis

We at ATR have been working for some time on corpus-based synthesis techniques, and have developed and labelled many large single-speaker corpora. This section presents a brief summary of our synthesis work, followed by a case study illustrating our experiences using ready-made or publicly available corpora, and a description of the types of corpus we have found useful.

Recent advances in speech synthesis technology have facilitated the re-use of very large speech corpora (of approximately an hour of speech) for high-quality 'personality-preserving' voice synthesis [1]. By directly concatenating raw waveform segments from the speech corpora, without recourse to signal processing for modification of their prosody, quite realistic-sounding speech can be created but 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.

To date we have collected 91 such speech corpora for synthesis, from the voices of 55 different speakers, in five 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 monologue and conversation. Durations of recordings from a single speaker typically vary between twenty minutes and four hours, and we are currently analysing a sixteen-hour corpus from one speaker.

2.1 Random-access segment replay

Our system 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 recourse to subsequent signal processing (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 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. From these features we can 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 synthesis showing three emotional attitudes (sadness, anger, and joy) from a single speaker can be found at [3].

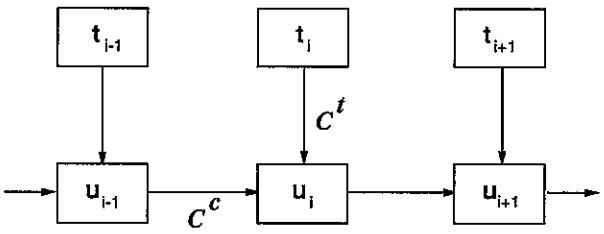

Figure 3: Two functions to select the speech segment that is closest to the target prosody while concatenating smoothly with its neighbours.

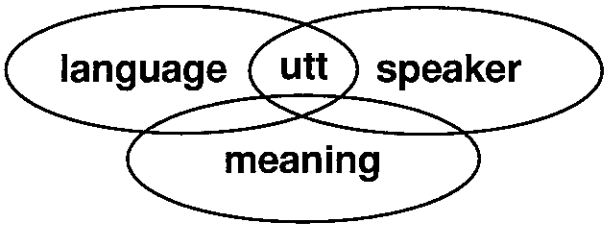

Figure 4: Three aspects of a spoken utterance

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 hidden Markov techniques. If only speech data is available, then we estimate a week of human labelling time per half-hour of recording. Once the 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 prepared.

From this full index, we select candidate units and find the best sequence for concatenation by Viterbi alignment, according to the two criteria illustrated in Figure 3. An optimal sequence of speech waveform segments will match the required prosodic targets for the utterance to be synthesised, while at the same time fitting smoothly enough together for the discontinuities at the joins to be imperceptible.

2.3 Three aspects of an utterance

Any given utterance has at least three relevant defining characteristics for such synthesis: the speaker, the language, and the intended meaning (Figure 4). These may be freely interchanged. For example, two people saying 'Hello' are probably using the same language with the same intended meaning, but if one were to say 'Bonjour' instead, then only one dimension would be changed (there is no requirement in the latter case that the speaker should

phonetic labels of the waveform, and the digitized waveform itself. The three kinds of information are stored in files for each category: 'txt' files are text files of orthographic transcriptions, 'lab' files are text files for phonetic labels, and 'wfm' files are binary files containing the digitized waveforms. Each individual utterance is stored in separate 'wfm' file. All the orthographic transcriptions and the phonetic labels of the utterances spoken by each speaker are grouped into the two related files, 'txt' and 'lab'.

We selected two speakers' speech data for testing (a male (4m) and a female (9f)), and used only the continuous speech for Chinese synthesis within CHATR. In the cs4m database, there are 975 utterances (about 45 minutes of speech), and in the cs9f database, there are 973 utterances (about one hour of speech). The two speakers selected represented the longest samples in the corpus, but as can be seen ([6]), while 9f can be considered adequate in length for synthesis purposes, 4m is probably not.

3.1.1 Aligning the corpus data

The phonetic labels in the distributed HKU database distinguish all the initial parts and some semi-triphones. The alignment of the phonetic labels is approximate and was done by auto-alignment. It is therefore not ideal. The following shows the phoneme set as labelled in the corpus.

- the initial parts (all consonants)
- the semi-triphones:

a	a<i>			
o	o<u>			
e	<i>e	e<i>	e<n>	e<ng>
<ch>i	<c>i	<b>i		
u	yv	er	ng	

When we used the original labels to make a test database for synthesis, we found the results to be unsatisfactory, mainly due to the following reasons: (a) Chinese is a tonal language but the tone information was missing from the label files, (b) about 50% of the phonetic labels were inadequately aligned, and (c) the same phonemic label ('n') was used for both the initial parts and the final parts although these sounds are different in Chinese. Similar problems have been encountered with every distributed database we have so far processed.

Because Chinese syllables can be defined by the combination of their initial and final parts, with each syllable thus formed having one of 5 possible tones, we re-categorised the initials and finals into distinct phonemic labels. The tone information was then marked explicitly (by rule from the orthography) onto the labels of the finals. In this way, phonemically similar final parts with different tones were relabelled as different phonemes. This resulted in a total of 220 phone label types for the Mandarin speech data.

Since the two selected databases were already segmented and labelled in the distribution version of the corpus, we automatically converted the original label files to the format defined by the new phone set and then manually corrected the units which had shown problems in alignment.

3.1.2 Language-specific unit-selection

The strategy of unit selection is based on minimizing both (a) distance between target segment and selected unit, and (b) distance between selected unit and previous selected unit, i.e. the join or continuity distance. In theory, with an infinite database, the actual unit selection is not a problem, because there will be enough units to choose from in any given situation. But few existing corpora are big enough to approximate this theoretical state. It happens quite often that there's no unit in the database which fulfills both the segmental and the prosodic requirements that have been predicted. In this case perceptually equivalent units have to be found. This process forms the core of the CHATR algorithm.

For Chinese, the syllable initial parts are all consonantal, but the characteristics of each vary according to the final parts, which specify the vocalic features. In our relabelling of the database, we used the same phone label to represent a final part whether or not it was preceded by an initial. For example, to synthesize the syllable "qie1", we need a unit "q" and a unit "ie1". If the consonant unit "q" is selected from a context like "qin1", then the resulting syllable will be good because "in1" and "ie1" have similar initial acoustic features, but if it comes from a context such as "qu1", then the resulting speech will be imperfect because "yv1" and "ie1" have quite different features. To avoid such inadequate selection, we need to distinguish the equivalent unit classes automatically. Final parts which begin with "i", "u" or "ü" have pronunciations which vary significantly, but for the synthesis we can distinguish such cases according to the preceding unit label (vowel, consonant or silent pause).

Examples of the resulting synthesised speech in Chinese can be found at [6], and will be presented at the workshop. To date we have performed similar processing with the commercially available Kiel Phondat CD-Roms (with permission), and with in-house recordings of Korean and Japanese as well as with British and American English.

3.1.3 Text types for synthesis

Although the HKU corpus contained samples of many text types (including Isolated Syllables, Words, Digit Strings, Rhymed Syllables, Continuous Speech, and Retroflexed Ending Words), only the continuous speech was considered to be of use for our synthesis purposes.

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

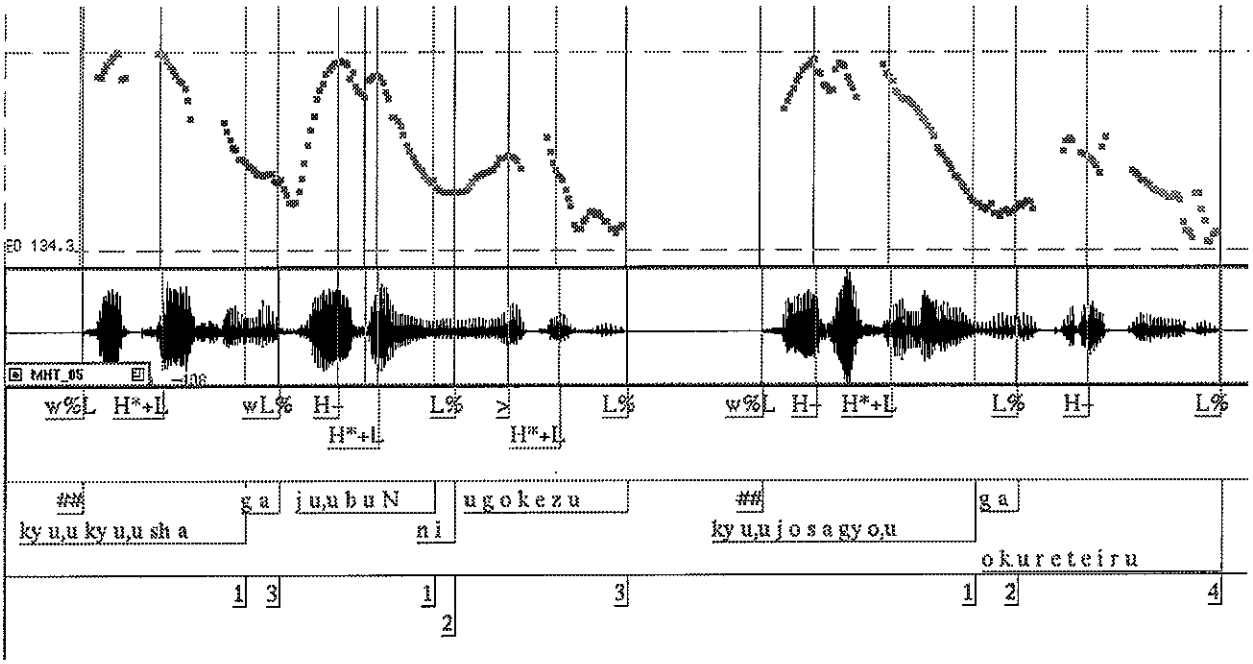

Figure 2: Labelling a speech corpus for phonemic and prosodic information about each segment.

be different). Similarly, a speaker saying 'hello' on two different occasions may be performing two different functions (for example, greeting vs exclamation), thus varying meaning. While such 'meaning' differences may be the hardest to control for synthesis, we can easily interchange the other two parameters, and for our interpreted telecommunications research we are particularly interested in using the voice of one speaker to reproduce 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 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 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 produces a given utterance, then we can use this 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 such as is required for speech translation.

In his section, we will address the issue of producing the original (target) vector sequence using an existing corpus of speech in a foreign language.

3.1 Chinese Speech Synthesis

This case study reports on work carried out by a visiting researcher at our lab, adapting the HKU Mandarin Corpus (CD-ROM Version) [4] for use with CHATR [5]. This database was constructed at the Speech Laboratory of the Department of Computer Science, at the University of Hong Kong in 1993-96. A total of 20 native Mandarin speakers were employed to read prompt messages displayed on a monitor screen. The data are stored in five CD-ROMs.

The HKU Mandarin corpus contains the following speech types:

- **Isolated Syllables:** all Mandarin syllables in all tones
- **Words:** 11 words of 2 to 4 syllables are selected in such a way that their pronunciations include all the Mandarin phones.
- **Digit Strings:** 16 digit strings of 4 to 7 digits each are designed to exhaust all inter-digit triphones.
- **Rhymed Syllables:** 3 sentences of 7 syllables each are so designed that all syllables in the first sentence rhyme with /a/, those in second rhyme with /i/ and those in third rhyme with /u/.
- **Continuous Speech:** several hundred lines of text with unique contents.
- **Retroflexed Ending Words:** A set of words with and without retroflexed endings have been read by two speakers.

The information on the CD-Roms also includes the orthographic transcription (Chinese characters in Big5 code) with the toned Pin-yin symbols, the

Subsequently, 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 diffficult for most of our speakers to pronounce (having been inserted to ensure full coverage of all triphone combinations), we found that the tension (or lack of interest) during the recording session remained in the voice and resulted in '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 resulting phonemic distributions 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.

4 Multi-lingual Synthesis

An interesting and unexpected application of the use of multilingual corpora is in the area of cross-language synthesis. Originally, in order to reproduce the speech of one person with the voice of a speaker of another language, we used text-based mapping vectors to convert between a phone label in one language and the label identifying the equivalent sound in a target language. However, this resulted in heavily accented synbthesised speech (see for example [8]), But if there is enough similarity between the voice types of the source- and target-language speakers, then the cepstral-target method of unit selection currently being developed offers intonation closer to that of a native-speaker.

For example, in the case of English speech from a voice synthesised using the waveforms of a Japanese speaker, the label information alone is not adequate to distinguish the different (RP) vowel sounds in the words 'cap' and 'cup' (both being mapped onto the same Japanese vowel /a/). However, in the speech of most Japanese, there is sufficient variation within the pronunciation of /a/ tokens to include sufficiently representative versions for each required English vowel sound. Similar variation in the speech can be found for the /l/,/r/ pair which are not phonemically distinguished in Japanese, and are therefore (in spite of production differences) usually marked with the same label. By using spectral information as a direct target in the waveform unit selection, we are able to reduce the ambiguity of the provided labels and to generate more intellligible synthetic speech in the 'foreign' language.

For unlimited use of this synthesis method as a text-to-speech conversion device, a language processing component for each target language would also be required, in order to convert the orthography of a target text into its phonemic representation and corresponding appropriate prosody. However, there are many applications of speech synthesis that do not require raw text as input, and many are of the opinion that adequate interpretation of the meaning of an input text is still beyond the capabilities of machine processing if 'natural' and expressive intonation is required. Synthesis from generated text, however, is another matter.

5 Copyright issues

I have been informed (and would definitely like to hear if there are contrary opinions) that there is no copyright on a speaker's voice. Parallels are drawn with the colours in a painting or the individual words in a book. Since it is only the original combination of these basic units that can be subject to copyright, the basic units themselves are considered to be in the public domain. We can expect changes in the legal situation if such synthesis methods as CHATR become more widespread, but for the immediate future, care must be taken that 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 mistaken to be speech spoken by the original speaker then they might cause 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.

Conclusion

This paper has presented examples of the multilingual application and reuse of existing corpora for synthesis both in the original language and across languages. Speech data for all the above examples of large-corpus-based speech synthesis can be found at www.itl.atr.co.jp/chatr along with many other examples showing the potential of this approach.

We pointed out in the discussion of this system that linguistic features alone are not sufficient to categorise the meaningful variations in speech and that such environmental factors as speaker's emotional attitude and quality of voice phonation are meaningful attributes that should be also be included in any 'phonetic' design of a speech database.

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 COCOSDA is the most appropriate forum in which to make this plea to corpus developers so that when future data collections are planned, there be allocated at least one speaker of each sex who will speak for sufficient time to allow enough prosodic and phonemic variation.

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 an automatic speech translation algorithm.

Acknowledgements

I would like to express particular thanks to Professors Chorkin Chan of Hong Kong University, and Klaus Kohler of Kiel University, Germany, for making their speech corpora available for CHATR research and experimentation, and to Ming Yue Xie-Zhang for assistance with its development.

References

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

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

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

[4] Y.Q.Zu, W.X.Li, M.C.Ho, C.Chan *HKU96 – A Mandarin Corpus CD-ROM Version.* Speech Lab. Dept. of Computer Science Univ. of Hong Kong, 1996

[5] ATR Tech Rept TR-IT-0243 "Chinese Speech Synthesis within Chatr", Ming Yue Xie-Zhang, 1997.

[6] http://www.itl. atr.co.jp/chatr/chinese.html

[7] ATR Tech Rept TR-IT-0236 "German in eight weeks - a crash course for Chatr, Caren Brinckman, 1997.

[8] http://www.itl.atr.co.jp/chatr/german.html