A call for generic-use large-scale single-speaker speech corpora and an example of their application in concatenative speech synthesis

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 how large-scale single-speaker speech corpora can be used for speech synthesis, and argues for the inclusion of more single-speaker data in the speech corpora that are currently being planned and collected in various parts of the world. Most of the large speech corpora currently available were designed for the training of speech recognisers and include many short speech samples from a wide variety of speakers. Such data are useful for the development of speaker-independent recognition systems, but the speech samples are typically limited in duration to the order of a few minutes each, and therefore offer little of interest to the synthesis, discourse, and prosodic analysis communities, whose needs are for similar but longer samples of speech for the analysis of speaking style, paragraph-level prosody, and discourse structure. I argue here that consideration of such needs would place only small demands on the design and collection of future speech corpora, and could be met by extending recordings for even just one speaker in the corpus. As an example of the wider application of such data we describe the processing of one such corpus for a speech synthesis system that makes use of external corpora of single-speaker data for source units.

1 Introduction

An earlier version of this paper was presented at the first Oriental-COCOSDA meeting in Tsukuba in May 1997. COCOSDA was established to encourage and promote international interaction and cooperation in the foundation areas of spoken language processing, and emphasises the importance of collaboration which transcends national boundaries. It is therefore an ideal forum at which inter-disciplinary requirements should be discussed. However, as COCOSDA is currently organised, the three separate working groups (Recognition, Synthesis, and Labelling) encourage little interaction between the different communities, which are perceived as having non-overlapping needs, and the means for making cross-disciplinary requirements more widely understood are still not yet well enough established. The present paper therefore argues also for a form of collaboration which also transcends the disciplinary boundaries.

Many large corpora of speech are now being collected and distributed, but their design is dictated by the needs of the 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 from any single speaker. To maximise the usefulness of such corpora, we need to consider collecting materials that not only meet the present requirements but that may serve broader future needs as well. For example by introducing longer reading passages for one or two speakers and by including more conversational and spontaneous

speech materials, we can provide data for use in such disciplines as dialogue research, speech synthesis, and prosodic information processingm as well as laying the foundation for future speech recognition.

Since the collection and annotation of large speech corpora is an expensive and time-consuming process, we need to take extra care and consideration when designing and evaluating such materials to ensure that they meet the needs of the widest possible range of scientific and industrial applications. In many cases, only minor changes to the corpus design would be necessary, and need not be required for all speakers of the corpus. Considering the expense involved in the original design and collection of a speech corpus, these small changes need not involve significant extra cost.

2 Synthesis from Speech Corpora

This section presents a brief summary of our corpus-based synthesis, followed by a case study illustrating some experiences using ready-made or publicly available corpora, and a description of the types of corpus we have found useful.

2.1 Corpus-based concatenative synthesis

At ATR we have been working with corpus-based speech synthesis techniques for some time [1, 2, 3]. Recent advances in speech synthesis have enabled the re-use of large speech corpora (containing approximately an hour of speech) for high-quality 'personality-preserving' voice synthesis [4, 5, 6]. By re-sequencing, or directly concatenating raw waveform segments from an external speech corpus without recourse to signal processing for modification of the prosody, human-sounding synthetic speech can be created, but the cost of such synthesis is that the source corpus must be big enough to contain multiple examples of all the basic speech sounds in each typical prosodic context for every candidate speaker.

To date we have collected 97 such speech corpora for re-sequencing synthesis, using 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, though we are also currently analysing a sixteen-hour corpus from one speaker.

2.2 Variability in a single-speaker speech corpus

FIGURE 1 ABOUT HERE

CHATR's system of indexing speech segments according to their joint phonological and prosodic attributes allows the selection of candidate waveform segments with sufficient accuracy for concatenation without recourse to subsequent signal processing (as illustrated in Figure 1). By specifying the variability in these two dimensions we are able to characterise the speech of any given speaker, but only for the given style of speaking in the corpus. However, 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 variation in emotional attitudes increases, with consequent discontinuities becoming noticeable 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 differences in voice quality (phonation) to show emotion. Examples of synthesis distinguishing three emotional types (sadness, anger, and joy) from a single speaker can be found at [10].

2.3 Text types for a synthesis corpus

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 word lists is not at all representative of the intonation required for continuous speech, and the resulting sounds themselves tended to be over-articulated, presumably to emphasise the contrasts between the words in absence of a defining textual context.

Consequently, in an intermediate stage of CHATR development, we tested speech from readings of a set of 500 phonemically-balanced sentences for use as source units for the concatenative synthesis. However, as these texts contained many sound-sequences, or phone combinations, 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 these 'random' corpora (Kuroyanagi Tetsuko reading for an hour, n=30,353 units) has shown little diffference between the resulting phone distributions and those of more carefully designed corpora (ATR 503 sentences, n=29,231 units), once a certain size threshold has been reached. However, the relaxed state and 'interested' tone-of-voice that arises from reading of meaningful continuous text results in a pleasant and natural voice quality and provides representative prosodic variants in the corpus.

2.4 Labelling a speech database

FIGURE 2 ABOUT HERE

An example of the phonological and prosodic labelling is shown in Figure 2 (segment labels omitted for clarity). 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 produced by rule and aligned using hidden Markov techniques. If only the 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 autommatically determined and a full index into the corpus produced.

FIGURE 3 ABOUT HERE

From this full index, we select suitable candidate units to match a target utterance for synthesis and find the best sequence for concatenation by Viterbi alignment, according to the two criteria (target cost and join cost [6]) 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.

2.5 Three aspects of an utterance

FIGURE 4 ABOUT HERE

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 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, speaker and language, 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 Processing an Existing Corpus for CHATR Synthesis

In this section, we address the issue of processing an existing speech corpus for the synthesis of speech in a foreign language, so that the voice of the input speaker can be used to produce utterances in the (translated) output language.

3.1 The HKU Mandarin Corpus

This case study reports on work carried out by a visiting researcher at our lab, adapting the HKU (Hong Kong University) Mandarin Corpus (CD-ROM Version) [11] for use with CHATR [12]. 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 CD-Roms include the orthographic transcription (Chinese characters in Big5 code) and their equivalent tone-marked Pin-yin symbols, the phonetic labels for each utterance, and the digitized speech waveforms. 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'.

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 selected two speakers' speech data for testing (a male (cs4m) and a female (cs9f)), and used only the continuous speech for Chinese synthesis within CHATR. In the male speaker's database, there are 975 utterances (about 45 minutes of speech), and in the female speaker's 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 ([13]), while the female speech database can be considered adequate in length for synthesis purposes, the male one is probably not.

3.2 Re-aligning the corpus data

The phonetic labels in the distributed HKU database distinguish the syllable onsets and rymes (all vowels, diphthongs, and any nasal coda segments). The alignment of the phonetic labels is approximate and was done by auto-alignment. The following shows the phoneme set as labelled in the corpus. Angle brackets indicate segment sequences (such as diphones) that have strong co-articulation effects and are difficult to divide into separate segments.

- the onsets (all consonants)
- the rymes:

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 onset and ryme segments although these sounds are phonetically different in Chinese (as in Japanese). Similar problems have been encountered with the suitability of supplied labelling in every distributed database we have processed. Because Chinese syllables can be defined by a combination of the onset and ryme segments, with each syllable thus formed having one of 5 possible tones, we re-categorised the ryme segments into distinct phonemic types and the tone information was then marked explicitly (by rule from the Pinyin orthography) onto the labels of the vowels. In this way, phonemically similar rymes with different tones were explicitly relabelled as different types. 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 trained new HMM models for re-alignment, and finally manually corrected the units which had shown problems in alignment.

3.3 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 weight-training algorithm.

In our relabelling of the database, we used the same phone label to represent a syllable ryme whether or not it was preceded by an initial consonant. 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 satisfactory 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 noisy because "yv1" and "ie1"

have quite different acoustic features. To avoid such inadequate selection, we need to distinguish the equivalent unit classes automatically. Syllable rymes which include "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 [13]. 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.

4 Cross-language Synthesis

An interesting, if unexpected, application of the use of multilingual corpora is in the area of cross-language synthesis. Originally, in order to reproduce speech in one language using the voice of a speaker of another language, we used text-based mapping vectors to convert between a phone label in the source language and the label identifying the equivalent sound in the target language. However, this resulted in heavily accented synthesised speech (see for example [16]), 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 [9] offers intonation and segmental quality closer to that of a native-speaker.

4.1 Voice mapping across languages

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 phonemic and prosodic 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 reproduce the utterance as it would be 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 a speech translation system.

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 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 allophonic variation can be found for the /l/,/r/ pair which are not phonemically distinguished in Japanese, and are therefore (in spite of phonetic differences) usually marked with the same phonemic label. By using spectral information as a direct target in place of abstract feature sets in the waveform unit selection, we are able to reduce the ambiguity of the corpus labels and to generate more intellligible synthetic speech in the 'foreign' language.

4.2 Linguistic processing and input specification

For unlimited use of this synthesis method as a text-to-speech conversion device, a language processing component for each target language is also 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 it can be argued 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.

Speech samples illustrating the above examples of large-corpus-based speech synthesis can be found at www.itl.atr.co.jp/chatr along with other examples showing the potential of this approach.

5 Copyright issues

I have been informed [15] (and would definitely like to hear if there are contrary opinions) that there is no copyright on a speaker's voice per se. 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's rights (moral as well as legal) 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 actual 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.

6 Conclusion

This paper has presented examples of the multi-lingual application and re-use of existing corpora for synthesis both in the original language and across different languages. It was pointed out in the discussion of this system that linguistic features alone are not sufficient to categorise the range of meaningful variations in speech and that such contextual factors as speaker's emotional attitude and quality of voice phonation are also attributes that might be considered in the 'phonetic' design of a future speech database.

It would be of great interest to develop this technology further, using speech data from other languages and with 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 this is an appropriate place to make a 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 potential 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 process.

Acknowledgements

I would like to express particular gratitude 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 and Caren Brinckman for their assistance with adapting the corpora for synthesis.

References

[1] Y. Sagisaka (1988), "Speech synthesis by rule using an optimal selection of nonuniform synthesis units", Proc. IEEE-ICASSP88, 679-682.

[2] Y. Sagisaka, N. Kaiki, N. Iwahashi, and K. Mimura. "ATR ν-talk speech synthesis system". In *Proc. 1992 Intl. Conf. on Spoken Language Processing*, pages 483–486, Banff, Canada, 1992.

[3] Y. Sagisaka and N. Iwahashi. "Objective optimization in algorithms for text-to-speech synthesis". In *Speech Coding and Synthesis,* W. B. Klein & K. K. Paliwal, Eds., Elsevier Science B. V. 1995.

[4] W. N. Campbell. "Synthesis units for natural English speech". Technical Report SP 91-129, IEICE, 1992.

[5] W. N. Campbell. "Prosody and the selection of source units for concatenative synthesis". In Proc 2nd ESCA Workshop on Speech Synthesis, Mohonk, N.Y., 1994.

[6] W. N. Campbell and A. W. Black, "CHATR: a multi-lingual speech re-sequencing synthesis system", 45-52, SP96-7 Tech Rept IEICE, (Japanese) 1996(5).

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

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

[9] W. N. Campbell, "Multi-lingual concatenative speech synthesis",Proc ICSLP-98.

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

[11] 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

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

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

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

[15] New York Times (Business Section) April 21st 1997.

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

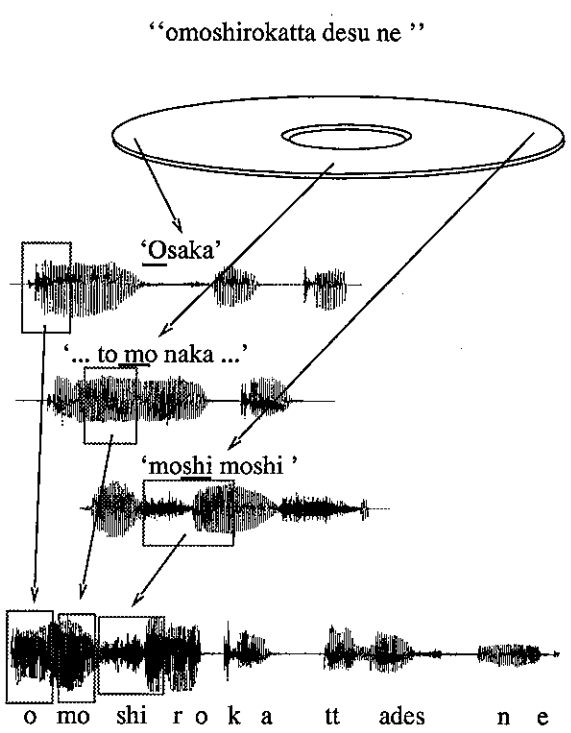

Figure 1: Selecting speech waveform segments to create novel utterances using the voice of the original speaker. The target utterance 'omoshirokattadesune' (that was interesting) is rendered by concatenating short non-uniform-length waveform samples selected according to prosodic and phonemic environment from amongst the candidates available in the corpus.

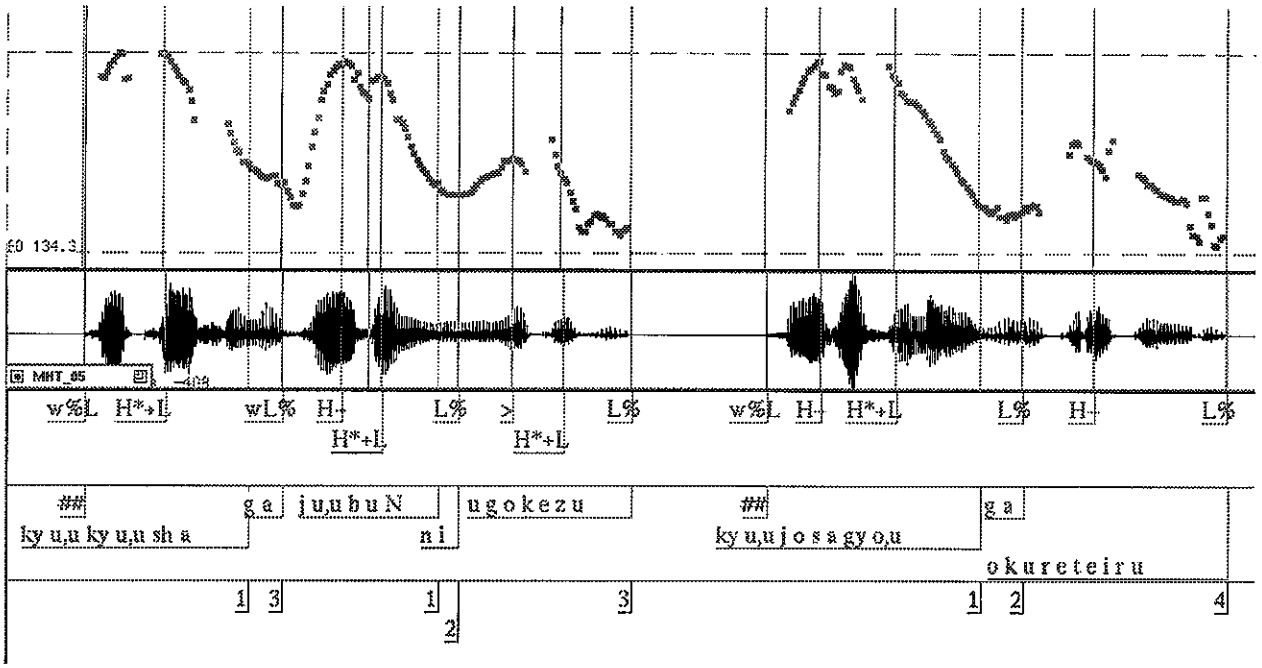

Figure 2: Labelling a speech corpus for phonemic and prosodic information about each segment. Shows F0, speech signal, and ToBI labels. The three tiers visible beneath the speech wave represent the tones, the word sequence, and the break indices respectively.

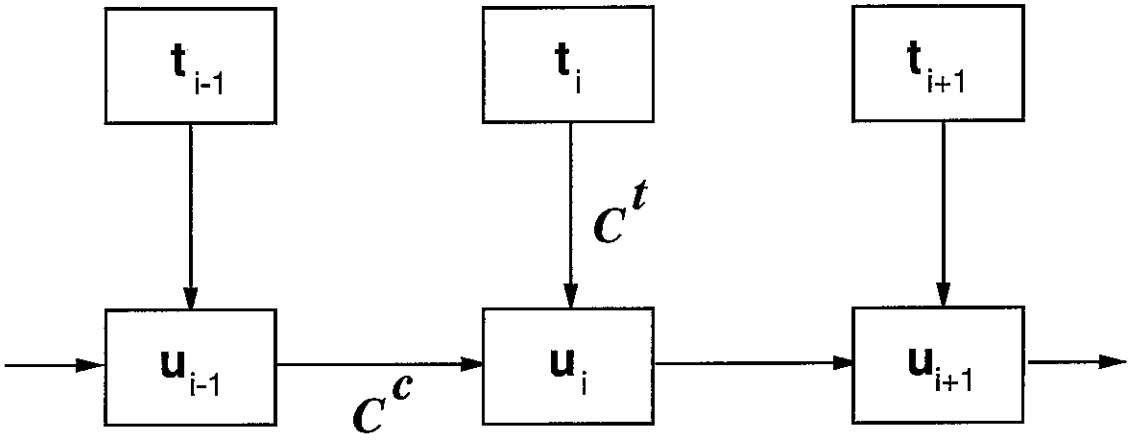

Figure 3: Two functions to select the speech segment that is closest to the target prosody while at the same time joining smoothly with its neighbours. Weights C^c and C^t are trained so that the two costs can be minimised simultaneously to find an optimal sequence of units for concatenation from the corpus. Here t-1, t, t+1 represents the target sequence of ideal phonemes for synthesis, and u-1, u, u+1 represents the candidate segment sequence available for concatenation from the current database

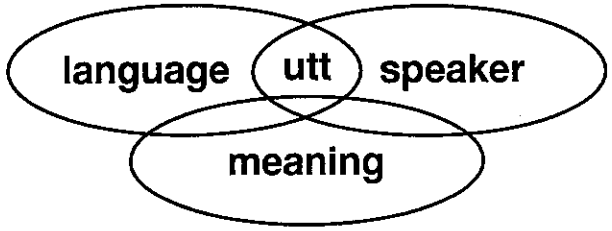

Figure 4: Three aspects of a spoken utterance (utt): Language can be represented through the sequencing of the speech sounds, meaning expressed through the prosody assigned to the utterance, and the speaker characteristics taken directly from the database.