Factors to Consider in the Design of an Optimal Speech Corpus for Concatenative Speech Synthesis. *

◎ Nick Campbell & Ekaterina Saenko
ATR Interpreting Telecommunications Research Labs

Abstract

This paper describes factors that should be taken into consideration when preparing a speech corpus for CHATR synthesis, where in addition to phonemic balance, prosodic and phonation balance are also considered necessary. The paper describes a formalism that reduces the complexity of combinations of these factors and a system for generating and recording the texts automatically.

1 Introduction

Currently, CHATR databases [1, 2] rely on size to ensure a representative coverage of the speech sounds of a given language. However, given the large text corpora that are increasingly becoming available, we can now reliably predict the balance requirements (both phonemic and prosodic) for an optimal speech database design automatically.

Since the range of speaking styles is great, CHATR synthesis is more effective when domain-specific; for example, a source database of news readings will not necessarily be suitable for the generation of 'spontaneous' conversational speech, but may be adequate for car-navigation tasks. (see for example the audio web-page [3]).

For any given text corpus, which must be large enough to represent the domain it covers, synthesiser modules predict the bi-phonemic and prosodic characteristics, and the resulting over-specified description can be reduced to a small database of source texts for reading. Experience with CHATR has shown that more natural voice-quality results from reading continuous texts than isolated sentences, so our task is to produce representative paragraphs.

2 Text Generation

Details of the method for deriving a balanced text corpus are presented in [4] and will be summarised only briefly here. The commonly-used N-gram phone model assumes statistical dependencies between symbols of fixed length, N. However, in speech, some variable-length sequences of phonemes are equally likely, such that for example "pau-a-n-d", "a-n", and "o-v-dh-@" all have approximately the same likelihood.

The multigram model [5] considers a string of symbols as the concatenation of independent variable-length sequences. When applied to speech synthesis unit generation, it produces multiphone units based on their maximum likelihood. The model extracts variable-length regularities which are strongly related to the morphological structure of speech. Experiments have shown that on average, a sentence synthesized using multiphones has only half the number of (potentially disruptive) concatenations compared to that of biphones.

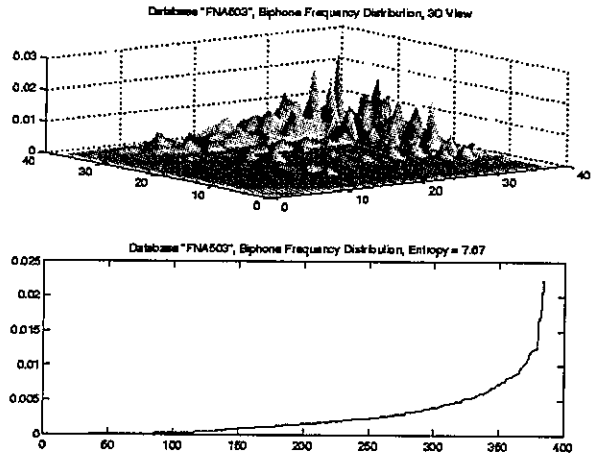

Fig. 1 Uneven distribution of bi-phones in a phonemically 'balanced' corpus (ATR503DB)

```
The smell of freshly ground coffee
never fails to entice me into the shop.
   pau-dh-@-s  m-e-l  o-v-f-r e-sh l-ii
   g-r-au-n-d k-o-f-ii pau-n-e-v-@ f-ei l-z
   t-@ e-n-t ai-s m ii-pau-i-n t-@-dh-@ sh-o-p

The chill wind caused them to shiver violently.
   pau-dh-@ ch-i-l w-i-n-d k-oo-z-d  dh-e-m
   t-@-sh i-v-@ v-ai l-@-n-t l-ii
```

As can be seen from table 1, the storage requirements for a multiphone set are approximately 1.5 times the size of an equivalent triphone set, but the number of unique units is greatly reduced. Overall, multiphones scale well compared to triphones, and they provide the longer and necessary sequences that

Table 1 Coverage of a database according to size of N-gram unit for a CHATR English speech corpus

Type of Unit	Total	50%	90%	Storage
Biphone (N=2)	1,729	106	528	1,056
Triphone (N=3)	35,862	871	6,451	19,353
M-phone (1≤N≤5)	13,744	1,443	8,460	29,924

*音声波形接続合成専用テキストのデータベース作成について (ニック キャンベル ・エカテリナ サエンコ) ATR 音声翻訳通信研究所

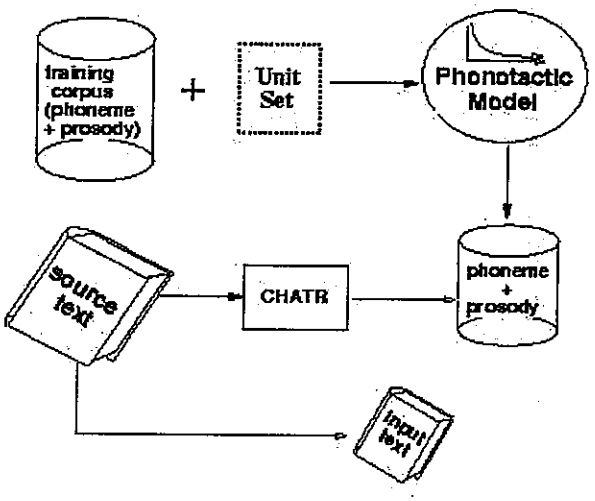

Fig. 2 Offline processing for a text-balanced speech database.

tend to be re-used in speech often, and which can be assumed to be more co-articulated and idiosyncratic.

By including a measure of prosodic boundary strength in the 'phoneme' set (coded as ToBI Break Index levels), the phonatory qualities of prosodic-phrase-initial and prosodic-phrase-final elements can be simply encoded in the 'phonemic' specification. It remains therefore to ensure balance in the accentual domain, which will be discussed in the next section.

3 Balancing the corpus

Because of the inclusion of prosodic characteristics at the unit selection stage of CHATR synthesis, it is important to ensure that the source corpus is balanced in terms of prosody as well as phonemic coverage, i.e., that it is 'representative' rather than 'uniform' (figure 1 illustrates a natural bias in the phonemic coverage of the ATR 503 sentence data), and we focus on balancing the prosodic aspects of segments to ensure that the resulting speech database is adequate for general-purpose synthesis within the domain.

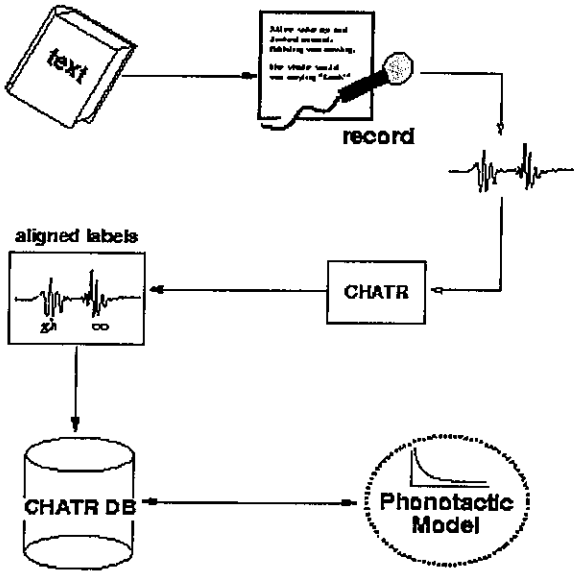

Fig. 3 Online processing for a prosodically-balanced speech database.

Before recording, the phonetic and prosodic properties of the database are predicted and the optimal balance is determine off-line from the pre-compiled input text (figure 2). However, the human interaction introduces new information and variability into the unit distribution. The sources of variability are, for example, idiosyncratic pronounciation of words (there are 80 pronounciation variants of the word "and" in the Switchboard corpus) insertion and omission of pauses, and presence or absence of stress, etc. It is therefore desirable to analyse the unit balance of the resulting waveform and compare it *during collection* to the originally intended model (figure 3). If there are significant deviations from the model, then we can re-insert some of the missed units by having the speaker interactively re-read selected parts of the text.

4 Collection sequence

Offline: A phonetic corpus is taken to be the "universe". The phonological model is trained on the corpus with respect to the unit set. Phonetics and prosody (pre/post boundary, +/- stress) are predicted for a source text using CHATR. The text is reduced using a balancing algorithm.

Online: The user is prompted with sentences from selected text. The user has control over the timing and editing of wave files. The system aligns the labels with the wave file after each text is recorded. Possible pronounciation variants and prosodic contours are predicted, choosing the best match. It then analyses and reduces the recorded units based on similarity measures determined for the phonotactic model. Finally, the distribution is compared with the original model and the loop repeated until the database is complete.

5 Conclusion

We have presented a method for preparing a speech corpus for CHATR synthesis, and a system for generating and recording the texts automatically. The examples given were for English, but the model can be used without adaptation for Japanese.

References

[1] W.N.Campbell, "Labelling an English speech database for prosody control", 1-P-8, Proc ASJ, Spring, 1992.

[2] W.N.Campbell, A.W.Black, "CHATR: 自然音声波形接続型任意音声合成システム", 信学技報,SP96-7 1996.

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

[4] Ekaterina Saenko & Nick Campbell, "Recording New Databases for CHATR", ATR TR-IT-0288.

[5] S. Deligne & F. Bimbot, "Language modeling by variable-length sequences: theoretical formulation and evaluation of multigrams", Proc ICASSP 95.