A Database Design for a TTS Synthesis System Using Lexical Diphones

Tanya Lambert¹, Andrew Breen²

¹School of Computing Sciences, University of East Anglia, Norwich, UK,
²Nuance Communications, Norwich, UK

Abstract
Database designs, if based on the premise that there are about 2000 diphones in English, as stated in many publications and on-line documents, are likely to render a database of diphones, which will fail to capture some important phonological phenomena of English. This paper proposes a TTS database, which is built from diphones inclusive of their syllabic stress; we term these units lexical diphones.

A comprehensive lexical diphone feature set is generated using a stress-annotated dictionary and continuous text and speech. A method based on multiple set cover algorithms, applied to wordlists of specialized English usage, and a knowledge-based phonological approach, are used to produce a core text corpus of 540 sentences. An objective evaluation of our database with other databases shows that our database (considering its size) has a higher concentration of lexical diphones; a subjective evaluation shows listeners’ preference for the speech where there are more lexical than phonemic units.

1 Introduction
Traditional database designs for speech recognition and synthesis systems consist of automatic or manual generation of a corpus of sentences from various literary genres according to phonetic or prosodic criteria [2, 3, 4, 5, 9]. The research by [4] considered building a large text corpus based on triphones with an objective of achieving a uniform coverage of the most frequent triphones. Database designs, which aim to cover the most frequent units, run a risk of performing badly when less frequent units are needed for synthesis. Optimal selection of text is generally achieved using some type of a greedy or genetic algorithm [2, 5, 4, 5, 9].

For open domain TTS synthesis systems, designing a speech database, which would contain all prosodic variations in English is a challenging task. It is implied in [1] that databases would have to be astronomical in size if all prosodic units are to be covered. Where the past research concentrated on improving the algorithm for text selection [1, 2, 9], not enough research has been done in terms of the database content. [3] emphasizes the need for a database corpus where sentences are “designed for synthesis as opposed to merely collected”.

Our database design uses three domains: phonetic dictionaries, continuous speech and text, and well documented and researched phonological phenomena in English as in [7,11]. Our approach is concerned with capturing producible and permissible phonological events in English speech (both of English and non-English origin (e.g. schnapps)).

The format of this paper is as follows: section 2 is an analysis of the distribution of lexical diphones and triphones in various text domains. A new database design and a set cover algorithm are outlined in section 3. Section 4 examines the coverage of lexical units in our database with Messiah, Arctic [3] and TIMIT-SI. A subjective evaluation is given in section 5, followed by our conclusions and further work in sections 6 and 7 respectively.

2 Lexical Diphones and Triphones
For a language such as English, where phonological and prosodic speech variations are related to the context’s stress, it is insufficient to consider units without associated accent patterns. We propose here a database design where units are defined by their lexical stress.

In what follows, we examine the feasibility of achieving a full-coverage of lexical diphones for 44 phone-based transcription for British English. Each phone in a word is given a stress label of its parent syllable. Three labels are differentiated, 1 if the syllable bears a primary stress, 2 if the syllable bears a secondary stress and 0 if it is unstressed. Initial diphones in words August, Augustine and Augustinian are the same phonemically, but the syllabic stress in these words clearly distinguishes their acoustic properties (e.g. /'æg:/, /ə'g/:, and /æg/ where ’ marks the primary and , marks the secondary stress). A phonetic analyzer tool was built for the purpose of extracting the relevant stress information from text sources.

We first examined the frequency distribution of lexical diphones and triphones in a comprehensive phonetic dictionary with greater than 300,000 orthographic entries and with extensive coverage of all morphological inflectional forms used in English. (Table 1).

<table>
<thead>
<tr>
<th>No of words</th>
<th>Unique lex. diph.</th>
<th>Unique lex. triph.</th>
<th>Unique phonemic diph.</th>
<th>Unique phonemic triph.</th>
</tr>
</thead>
<tbody>
<tr>
<td>dictionary</td>
<td>9210</td>
<td>88318</td>
<td>1670</td>
<td>26026</td>
</tr>
<tr>
<td>(≥300K words)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 1: Frequency distributions of lexical and phonemic diphones and triphones in an English phonetic dictionary.

If we assume that every word in the dictionary can be followed by any other word (this is an unlikely scenario in normal speech), we could obtain an overestimated figure of how many juncture diphones are producible in English. From the dictionary of approximately 300K words, we estimated that there are 15982 distinct producible lexical juncture diphones. However, not all of these combinations would be permissible by English grammar and syntax rules.
A more realistic account of the distribution of lexical juncture diphones in English can be calculated from various samples of continuous text and speech, retrieved from [6, 13], as shown in Table 2.

<table>
<thead>
<tr>
<th>Text Genre (No of Words)</th>
<th>Unique lex. diph. in text</th>
<th>Unique phonemic diph. in text</th>
<th>Unique lex. juncture diph.</th>
<th>Unique phonemic juncture diph.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Writings of A. Lincoln (~38000)</td>
<td>3146</td>
<td>1299</td>
<td>1826</td>
<td>845</td>
</tr>
<tr>
<td>A sample of spoken English (78148) [10]</td>
<td>4262</td>
<td>1454</td>
<td>2540</td>
<td>1028</td>
</tr>
<tr>
<td>The Mill on the Floss by G. Eliot (266290)</td>
<td>5544</td>
<td>1539</td>
<td>3189</td>
<td>1138</td>
</tr>
<tr>
<td>various text genres (1.75 million words)</td>
<td>8331</td>
<td>1635</td>
<td>5331</td>
<td>1331</td>
</tr>
</tbody>
</table>

Table 2: Distribution of lexical diphones in various text genres

Out of approximately 15982 producible lexical juncture diphones in English, only 33.4% were encountered in the collection of texts from various genres containing over 1.75 million words. The variability of lexical diphones at word boundaries in text and speech is much greater than the variability of lexical diphones within words. As juncture diphones combine phones from different syllables, phones within a juncture diphone may carry different stress.

Frequency distributions of lexical diphones and triphones are dependent on a text genre. Less than 50% of lexical diphones, and less than 20% of lexical triphones, may be common to different text genres, as shown in Table 3. With regard to similar genres, about 85% of lexical diphones and 74% of lexical triphones were found to be common to two novels by the same author (i.e. Pride and Prejudice and Sense and Sensibility by J Austen).

<table>
<thead>
<tr>
<th>Text genre (ca. 100K-120K words)</th>
<th>Lex. Diph. (unique)</th>
<th>Lex.Triph. (unique)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pride and Prejudice, J Austen</td>
<td>4404</td>
<td>28477</td>
</tr>
<tr>
<td>Newspaper / adverts</td>
<td>5275</td>
<td>33747</td>
</tr>
<tr>
<td>Science Monthly</td>
<td>5517</td>
<td>38135</td>
</tr>
<tr>
<td>Names and Surnames</td>
<td>6082</td>
<td>45328</td>
</tr>
<tr>
<td>Common to all above texts</td>
<td>2747</td>
<td>7250</td>
</tr>
</tbody>
</table>

Table 3: Frequency distribution of lexical diphones and triphones in specified text samples.

Text corpora with words > 100,000 can have about 20% of lexical diphones with hapax legomena frequency. In text samples with less than 1000 words, hapax legomena diphones can account for more than 60%, as illustrated in Figure 1.

3 New Speech Database Design

The first stage in our database design, which is shown in Figure 2, is concerned with a compilation of an extensive lexical diphone feature set from our three main text domains. Multiple set cover algorithms are applied to samples of continuous text and auxiliary text sources. Auxiliary text sources consist of names and surnames’ lists, wordlists of frequent and specialized English usage (e.g. newspaper and technical English). Extensive manual corrections are used to ensure that varied syntactic structures are represented in the text, and that difficult tongue-twisting phrases are made easier to pronounce.

Phonological phenomena [7,11] (nasal assimilation, elisions, consonant voicing and devoicing, clear and dark /l/’s, vowel reductions, linking-r and r-controlled vowels, juncture plosives, if not found during automatic selection, are manually added to the database. Similarly, the coverage of all diphone stress patterns in homographs and the diversity of monosyllabic, di-, tri- and polysyllabic words are represented.

3.1 Set Cover Problem

For a phonological feature (e.g. lexical diphones, triphones) \( P = \{ P_1, ..., P_n \} \) and a text source \( S = \{ S_1, ..., S_m \} \), the set cover problem [8] can be represented by a matrix \( A = \{ a_{ij} \} \) where the matrix element \( a_{ij} = 1 \) if \( P_i \subseteq S_j \) and \( a_{ij} = 0 \) if \( P_i \notin S_j \). The maximum coverage of phonological features for a specific text source can be given by the equation (1):

\[
C = \max \sum_{j=1}^{m} a_{ij} \quad (1)
\]

The overall objective here is to find a minimum number of phrases/sentences from a given text corpus, which have the maximum number of covered units from the phonological feature set.

4 Evaluation of Text Corpus

The distribution of lexical diphones and triphones in our database (540 sentences) is compared with Messiah, used by the BT’s Laureate’s TTS system (279 sentences), one of the
Arctic databases, used by the Festival Speech Synthesis System (1132 sentences) and TIMIT-SI (1890 sentences). Although, according to the number of words used in each database, the relationship is Messiah < Our database < Arctic < TIMIT-SI, as shown in the inset bar chart in Figure 3, the distribution of lexical diphones and triphones in these four databases does not follow the same trend, as Figure 3 shows.

Figure 3: Lexical diphones and triphones in four databases

Both databases were recorded by the main author of this paper at the university’s recording studio as 44 kHz, mono, linear speech files, and subsequently treated in the same manner with regard to speech annotation process and voice generation.

The availability of lexical diphones and triphones per sentence is greater in our database (A) than Messiah (B). For the entire set of 90 sentences the difference in a number of units present in two text databases was found to be statistically significant at 95% CI and 99% CI with likelihood ratio, LR [12], of 502.4, p-value <0.0000 for lexical diphones and LR of 718.2, p<0.0000 for lexical triphones. With regard to individual sentences, the difference in the lexical coverage within two
databases was found to be statistically significant at 95% CI in 47% of sentences for lexical diphones (chi-square exceeding 4.1 and p-value < 0.04) and in 60% of the sentences for lexical triphones (chi-square exceeding 4.7 and p-value < 0.03). The difference in the lexical coverage in two databases was found to be significant at 99% CI for 32% of lexical diphones and triphones.

4.2 Unit Selection and Lexical Units
For the entire set of 90 sentences, 61% of lexical units were selected by unit selection when speech database A was used, and 50% of lexical units were selected when speech database B was used, as shown in Figure 5. The number of lexical triphones in sentences synthesized from speech database A exceeds that from B by 13.2%. In sentences synthesized from speech database A, lexical units are more abundant in 73% of the test sentences, this figure being significant at 95% CI in 43% of the test sentences.

5 Perceptual Evaluation
12 listeners gave their overall opinion score for ease of listening of synthesized speech on a scale of 1 to 5, 1 being bad and 5 being excellent. All listeners were adult native English speakers with no known hearing problems. Sentences were randomized in order to conceal from the listeners the identity of the speech database used in synthesis.

<table>
<thead>
<tr>
<th>Evaluation Scores</th>
<th>Speech Database A</th>
<th>Speech Database B</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>76</td>
<td>361</td>
</tr>
<tr>
<td>3</td>
<td>339</td>
<td>369</td>
</tr>
<tr>
<td>4</td>
<td>499</td>
<td>217</td>
</tr>
<tr>
<td>5</td>
<td>166</td>
<td>32</td>
</tr>
<tr>
<td>TOTAL</td>
<td>1080</td>
<td>1080</td>
</tr>
</tbody>
</table>

Table 6: Listeners’ evaluation scores

The MOS for speech database A is 3.70, SDev 0.81 and for database B 2.74, SDev 0.98. The scores’ variance for A database of 0.66 and for B database of 0.96 based on 1079 d.f. yield an F-ratio = 1.46, which is significant at 95% and 99% CI. In 4% of the test sentences listeners’ MOS were marginally higher for speech generated from database B. In these sentences the number of lexical units employed by unit selection is either the same or marginally higher than for database A.

6 Conclusions
The phonological content of a speech database is the blueprint from which TTS concatenative synthesis systems generate their speech. This paper proposes diphones with lexical stress as units, which are more appropriate than phonemic diphones for capturing phonological variations in English. Concatenative units in synthesized speech are more acceptable to listeners if they originate from a suitable syllabic environment and carry appropriate accent patterns. Traditional methods in database designs, which consist of collecting sentences from text genres, usually end up in large text databases, which are costly to record. Our modular design with multiple set cover algorithms, applied to text sources with different narrative styles and extensive manual corrections, produced a core database that is smaller in size but better in unit coverage.

7 Further Work
Our core database has been extended in order to achieve a near-full coverage of phonological units. Whilst we do not aim to cover every prosodic unit in our text database design, we aim to include phonological features of speech that from a perceptual perspective, really matter.

8 Acknowledgements
We thank the Nuance staff, Ellis Breen for help with auto-annotations tools, and Barry Eggleton for useful hints on improving the settings on the university’s recording equipment.

9 References