Expressive Prosody for Unit-selection Speech Synthesis

Volker Strom, Robert Clark, Simon King

Centre for Speech Technology Research
The University of Edinburgh, Edinburgh, UK
vstrom@inf.ed.ac.uk

Abstract

Current unit selection speech synthesis voices cannot produce emphasis or interrogative contours because of a lack of the necessary prosodic variation in the recorded speech database. A method of recording script design is proposed which addresses this shortcoming. Appropriate components were added to the target cost function of the Festival *Multisyn* engine, and a perceptual evaluation showed a clear preference over the baseline system.

Index Terms: speech synthesis, unit selection, prosody, recording script design

1. Introduction

The Festival unit selection speech synthesis system, *Multisyn* [1], achieves highly natural synthetic speech by avoiding use of an explicit model of prosody in terms of F0 and duration. Instead, large amounts of speech are recorded, so that each diphone is available in a variety of prosodic contexts. Of course, this means that there is no user control over the prosody of the resulting speech.

Even when F0 and duration models are used in unit selection systems with large databases, and these models are learnt from speech as is [2], they still represent an "average prosody" sounding somewhat unnatural and monotonous, similar to that of diphone synthesis.

The Festival *Multisyn* engine did not previously model prosody at all, except for distinguishing sentence-internal from sentence-final phrase boundaries. This works surprisingly well, provided that the database speech style closely matches the required synthesis style, and in particular is good for read newspaper-style text. But, for generating prosody appropriate for conveying specific meanings, such as emphasis or the type of a question, some prosody control is essential (note that, unlike [3], we are not attempting to convey *emotional* content).

It remains an open question as to on what level "prosodic context" is best described, and how to design a suitable text corpus. In our approach, we focus on emphasis and boundary tones, represented on the symbolic level.

2. Text corpus design

Corpus design, in this context also known as "recording script design" or simply "text selection", aims for maximal coverage of diphones *in context*. Above and beyond completeness of coverage, instances of diphones in multiple text types and reading styles are also desirable. Contextual features may include a stress flag for each half phone, the presence or absence of a boundary of a syllable, word, phrase, or sentence, (three possible locations), or more abstract descriptions such as ToBI accent and boundary tones, pitch accents, nuclear accents, sentence mood etc. The context may also include the identity of neighbouring phones.

To select those sentences to be recorded, a large text corpus is searched in order to determine the set of existing diphones-in-context. A subset of sentences is selected which covers them all, and which is as small as possible. The main advantage of making the subset as small as possible is the reduced effort of manually correcting the automatic annotations.

2.1. Standard approaches

In our standard approach, "context" refers to just syllable and word boundaries, plus lexical stress (as a binary feature, although the lexicon distinguishes primary, secondary and tertiary stress). In a text corpus of 442k newspaper sentences, 11585 distinct types of diphones-in-context were found. A subset of 7k sentences was selected that covers all of them.

But even in this restricted definition of context, complete coverage of all *existing* diphones is almost impossible to achieve. Figure 1 shows how many distinct diphones-in-context are found in subsets of the 442k newspaper sentences, made by randomly choosing 1/2, 1/4, ... of the corpus. Extrapolating this curve suggests that even 10 million newspaper sentences will not be enough to cover all diphones-in-context. This true even for a definition of context that ignores intonational context entirely. It is questionable whether covering diphones in all syllable boundary contexts is more desirable for synthesis than covering all diphones in, for example, all boundary tone contexts.

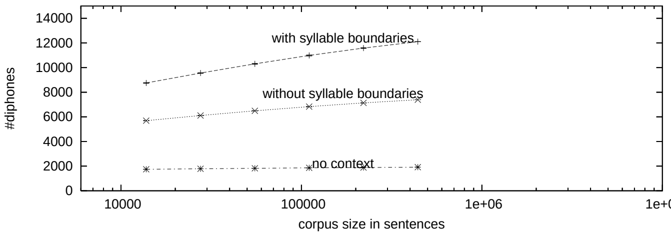

Figure 1: *Number of different diphones as a function of corpus size.*

Some approaches take more context into account, but still ignore prosody. [4] first covers all triphones found in a database of 153k sentences, then cover larger and larger units found in that database, up to morphemes with phone context. [5] attempts to cover as many pentaphones as possible.

In [8] and [3], small databases for specific prosodic expressions, such as contrast and yes/no-questions, were recorded in order to train speaker-independent prosodic models for duration and F0. For synthesis, a standard speech corpus is used (to save the cost of recording the specific prosodic expressions for each voice), hoping that it still yields what those prosodic models ask for.

Some *text selection* approaches aim to cover prosodic context in terms of phone duration and F0. [6] aims for an even distribution of the *predicted* F0 and duration, while in [7] the recorded speaker is explicitly told to read 525 sentences at three different speaking rates (slow, normal, fast) as well as in three pitch ranges (low, normal, high), yielding nine sub-corpora.

Models of expressive prosody are useless in unit selection synthesis if they ask for units that simply are missing from the speech database. This may sound trivial, but it is exactly the reason for the poor realisation of emphasized words in [8].

Selecting text on the basis of predicted prosody, on the other hand, is rather unreliable, due to the tendency of these predictors to average out the natural variance of prosody. Forcing a speaker to speak with a distinct pitch and rate, as in [7], runs the risk of unnatural-sounding joins when units from different sub-corpora are joined.

2.2. Our proposed approach

In order to avoid modelling prosody on the acoustic level, text marked up on a more abstract level was desired. Our initial idea was to use text generated by a dialogue system, which comes marked up with features such as given/new and theme/rheme. However, even the most sophisticated such system available to us, capable of producing 67k different sentences (about bathroom design, [9]) did not produce sentences of sufficient variety to form the basis of a general-purpose corpus.

In the search for text that has a natural variety of prosody, Lewis Carroll's children's' stories "Alice in Wonderland" and "Through the Looking Glass" seemed to be good candidates. Theatre plays and movie scripts were also considered, but the most alluring feature of Carroll's stories was their existing markup of emphasis using typographical devices.

As shown in Table 1, the major part of the recording script consists of word lists, read to achieve four different prosodic contexts, as described in Section 3.2. Some specialist texts were added: spelling, digit strings, and addresses. Finally, a relatively small number of newspaper sentences were used in order to cover any remaining missing diphones.

3. Corpus description

3.1. Carroll

From Lewis Carroll's children stories "Alice in Wonderland" and "Through the Looking Glass", 1434 sentences were selected manually from dialogue-rich sections. These are rich in questions, exclamations, quotations within spoken utterances and emphasized words such as contrastive and deictic pronouns. Furthermore, emphasized words are already capitalized in the text or are quotations within a spoken utterance:

...and even Stigand, the patriotic archbishop of Canterbury, found it advisable

'Found WHAT?' said the Duck.

'Found IT,' the Mouse replied rather crossly: 'of course you know what "it" means.'

'I know what "it" means well enough, when I find a thing,' said the Duck: 'it's generally a frog or a worm. The question is, what did the archbishop find?'

The speaker was asked to read in a spirited manner, but not to give the characters different voices.

3.2. Word lists

The largest part of the speech corpus consists of lists of 2880 words, selected from the Unisyn lexicon [10], such that all diphones (with context) in phrase-final syllable position are covered. Each word was read five times, with a fixed intonation pattern:

Ace, ace, ace. Ace? Ace!

Ache, ache, ache. Ache? Ache!

This covers continuation rise (L-H% at the commas), terminal intonation (L-L% at the period and exclamation mark) and interrogative intonation (H-H% at the question mark).

The speaker was asked to emphasize the last word. Thus, many diphones in emphasized words are covered, but by no means all of them, since the emphasis is mainly on the lexically stressed syllable, which in polysyllabic words is not necessarily the last one. The intonation was rehearsed at the beginning of each recoding session and was found to be fairly stable throughout the word lists sub-corpus.

Word selection criteria other than diphone coverage in word-final syllables were, in order: exclude homographs and function words, avoid proper nouns, prefer short words, and prefer more frequent words.

The main problem with this sub-corpus seems to be the poor performance of the automatic phone alignment method [1]. Because the word lists are not continuous speech like the other parts of the corpus, they appear to cause problems when training HMMs from a flat start. The resulting models have problems in accurately placing silence/speech and speech/silence boundaries. This affects not only stops, but also voiceless fricatives and sometimes even vowels. A number of improvements to the procedure described in [1] were attempted, but no satisfying solution has been found yet.

3.3. Newspaper

In the 442k newspaper sentences mentioned in Section 2.1, 6998 distinct diphones-in-context, but without the syllable boundary feature, were found. This set was compared to the diphone set covered thus far by the word list and Carroll's children stories. The difference was 2278 diphones, most being word-initial or across a word boundary. The fall-back strategy of the unit selection algorithm can fix these by inserting a short pause. But for the remaining 413 word-internal diphones, 283 newspaper sentences were selected to be added to the recording script.

3.4. Corpus statistics

Table 1 shows the sizes of the sub-corpora making up the entire speech database.

	sentences	words	minutes	
carroll	1400	10555	114	(27%)
wordlist	2880	14400	256	(61%)
address	45	401	3	(1%)
spelling	33	165	3	(1%)
news	282	6910	43	(10%)
total	4631	40916	419	(7h)

Table 1: Partitions of the speech database.

The speaking rate in words per minute varies considerably between sub-corpora: 56 for the word list, 93 for Carroll, and 161 for the newspaper text. This is reflected in the phone duration statistics: Figure 2 shows the duration distribution of /aa/.

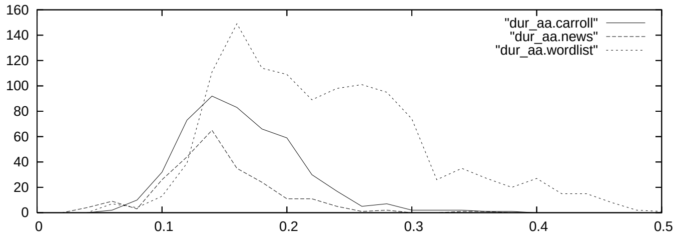

Figure 2: *Distribution of durations in seconds for the phone /aa/ in different subcorpora.*

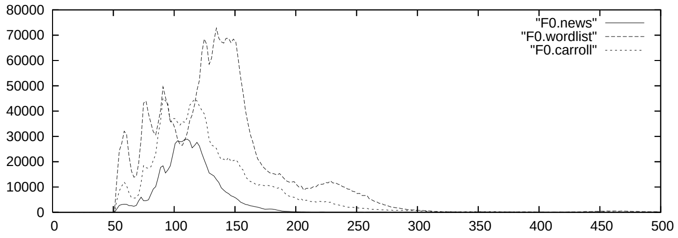

Figure 3: *Distribution of F0 values in Hz in different subcorpora.*

4. Automatic text-based prosodic labelling

The emphasis labels are solely based on textual markup: a word is considered emphasized if it is in uppercase, or if it is a short quotation (one or two words) within a spoken utterance, or is a short exclamation (including the 2880 such words in the word list sub-corpus).

In our system, the type boundary tone in ToBI notation is determined from punctuation, the POS of the sentence-initial word, and a flag indicating whether the sentence contains the word "or" immediately following a comma. Wh-questions get an L-L% boundary tone, yes/no questions an H-H%, and alternative questions (such as "1, 2, 3, or 4?") get an H-H% at each alternative but the last one, which gets an L-L%.

The sentence-initial word's POS distinguishes Wh-questions from other questions, although the interrogative pronoun may occur later, as in "But how do you know?" or "And what comes next?". The single worded question "What?" and its equivalents like "What was that again?" are also likely to be pronounced with an H-H%. An exception list is used to deal with these.

Questions which are not Wh-questions are either yes/no questions or alternative questions, which is decided by the "or" flag. This rule is imperfect: e.g. consider "Why can't we do that, or the people of Glasgow do that?", found in the newspaper text. The text corpus was manually checked for this type of Wh-question that looks like an alternative question. Recognizing them automatically is left as a future improvement; currently a user of the synthesis system is required to re-formulate the question as e.g. "Why can't we or the people of Glasgow do that?" in order to get the desired intonation.

Accents are not labelled or modelled yet.

5. Target cost function

The default target cost is a weighted sum of normalized components, which each score how well a candidate diphone matches the given target. These features are (most highly weighted first): lexical stress, phrase-finality, part of speech (noun, verb, or function word), position of the diphone in its syllable, position of the diphone in its word, left phonetic context and right phonetic context [1].

In the new system with prosody, a relatively large penalty is added when one half of the target diphone is a vowel and should be be emphasized, but the candidate is not, or vice versa. Another penalty is added when the boundary tones of target and candidate (L-L%, L-H%, H-H%, or NONE) do not agree. Note that there are no components for accent, F0, or duration. The relative weights for the target cost components were set manually and are not optimal.

6. Listening tests

Two web-based listening tests were used to evaluate the phrase boundary component and the emphasis component.

6.1. Boundary listening test

The purpose of this test was to find out whether the phrase boundary component improves the overall quality of the system, and *not* to test whether listeners recognize yes/no questions when they are not syntactically marked as such, as in [8].

Adding a target cost component for boundaries increases the pressure on the unit selection algorithm, because it exacerbates the problem of data sparsity. When the unit selection is pushed towards a particular intonation pattern at the cost of, for example, less smooth joins, the outcome may be worse than with the default system, even when the default system produced less appropriate intonation.

100 newspaper sentences were selected for the listening test, 25 each of: yes/no questions, Wh-questions, alternative questions, and statements. They ranged in length from 4 to 19 words with an average length of 9 words. They were synthesized by two versions of the Festival system: one using using all the target cost components described in the previous section, and one without the boundary cost component. The emphasis component was part of both systems, because this listening test is designed only to evaluate the boundary component. Without the emphasis component in the default system, emphasis can be realized somewhat arbitrarily, which would distract listeners from the main point of this test.

The order of the sentences was randomized, as well as the order of the two versions for each sentence. 10 volunteers, all native speakers of English, took the test. All but one listened with headphones. They were able to play each stimulus as often as they wanted, in any order, until they decided which version they preferred (this was a forced choice test with no "undecided" option).

The system with the boundary component was preferred in 557 of the 1000 stimuli-listener pairs (this is significant: $F(443; 1000, 0.5) < 0.0002$). Comparing the listeners, the number of votes for the system with the boundary component ranged from 48 to 60. For 32 of the 100 stimuli pairs, all or all but one listener agreed with each other.

6.2. Emphasis listening test

Sentence modality can be expressed syntactically, and even when a yes/no-questions does not start with a verb, speakers often do not raise the pitch at the end [11], because there is no need to do so if the nature of the question is obvious from the context. Thus there is no right or wrong question intonation: question intonation can only be more or less appropriate, formal, natural etc.

The purpose of emphasis, on the other hand, is much less "ornamental": it is often the only way to convey a special meaning such as contrast and is therefore essential, not optional. If speech synthesis system A succeeds in emphasizing the intended word in the perception of a listener, and system B fails to do so, it is not clear how to assess the segmental quality of both systems independently from system B's failure.

Therefore, in this second listening test, we looked at how well listeners recognize intended emphasis, regardless of other differences between the two systems (such as better or worse segmental quality). Seven short sentences were selected, 4 to 9 words long (6.3 on average), in which 3, 4, or 7 different words could carry emphasis. This resulted in 24 stimuli, the order of which was randomized.

Again, the volunteers could listen to each stimulus as often as they wanted and in any order. Their task was to locate in each sentence the one word which they perceived as most prominent. 15 listeners took the test, all but three of them with headphones. In 43% of the stimuli-listener pairs, the emphasized word was chosen correctly, with the average chance level being 18%. The average agreement between listeners was 83%.

	recog'd	agreem't	for
CAN he show us how to make it pay?	93.33%	93.33%	CAN
Can **HE** show us how to make it pay?	0.00%	73.33%	pay
Can he **SHOW** us how to make it pay?	20.00%	73.33%	pay
Can he show **US** how to make it pay?	0.00%	53.33%	pay
Can he show us **HOW** to make it pay?	80.00%	80.00%	HOW
Can he show us how to **MAKE** it pay?	0.00%	80.00%	pay
Can he show us how to make it **PAY**?	86.67%	86.67%	PAY
Average:	**40.00%**	**77.14%**	

Table 2: Recognition of emphasis shift.

Table 2 shows an example sentence before randomizing the stimuli order. It is the longest sentence, having 7 possible emphasis positions. Apparently it resulted in the most difficult set of stimuli: only in 40% of all 105 listener-stimuli pairs (7 variants $\times$ 15 listeners) the intended emphasis was recognized correctly. However, the chance level here is $1/7$, and that means the recognition rate is 2.8 times above the chance level.

7. Discussion

When it comes to phrase boundaries, in particular for yes/no-questions, our database should be big enough already. As mentioned at the end of Section 3.2, the major problem with this voice is still bad phone alignment, in particular at the boundaries between speech and silence.

[12] reports that although listeners prefer the standard intonation of yes/no-questions in natural speech, in synthesized speech their preference is based more on the overall quality than on intonation. It seems as if the best way to improve the listener judgement of the phrase boundary component is to further improve the phone alignment.

Looking closer at misrecognized emphasized words reveals that the recorded speaker did not put the same effort in all words marked up as emphasized. It is also striking that, if in doubt, listeners prefer the default location for nuclear accents, i.e. the rightmost pitch accents, as can be seen Table 2. However, when looking at where in the database the selected units come from, it becomes obvious that, as pointed out in section 3.2, many diphones in clearly emphasized words are still missing. Recording an additional 600 or so utterances in the word list sub-corpus style should close this gap.

8. Acknowledgements

The Author would like to thank Scottish Enterprise (under the Edinburgh-Stanford Link) for funding this project, and Roger Burroughes for his voice.

9. References

[1] Robert A.J. Clark and Korin Richmond and Simon King: "Multisyn voices from ARCTIC data for the Blizzard challenge", *Proc. Interspeech*, 2005

[2] V. Strom "From Text to Speech Without ToBI", *Proc. Int. Conf. on Spoken Language Processing*, 2002

[3] E. Eide, A. Aaron, R. Bakis, W. Hamza, M. Picheny, and J. Pitrelli: "A Corpus-Based Approach to <Ahem/> Expressive Speech Synthesis" *5th ISCA Speech Synthesis Workshop, Pittsburgh*, 2004

[4] M. Isogai, H. Mizuno and K. Mano: "Recording script design for corpus-based TTS system based on coverage of variuos phonetic elements from Speech Signals" *Proc. Int. Conf. on Acoustics, Speech and Signal Processing*, 2005

[5] B. Bozkurt and O. Ozturk and T. Dutoit: "Text design for TTS speech corpus building using a modified greedy selection" *Proc. Int. Conf. on Spoken Language Processing*, 2003

[6] H. Kawai, S.Yamamoto, and T. Shimizu: "A Design Method of Speech Corpus of Text-to-Speech", *Proc. Int. Conf. on Spoken Language Processing*, 2000

[7] H. Kawanami, T. Masuda, T. Toda and K. Shikano: "Designing Speech Database with Prosodic Variety for Expressive TTS System" *Proc. LRE*, 2000

[8] J.F. Pitrelli and E.M Eide: "Expressive Speech Synthesis Using American English ToBI: Questions and Contrasive Emphasis" *Proceedings ASRU*, 2003

[9] M. E. Foster, M. White, A. Stetzer and R. Catizone: "Multimodal Generation in the COMIC Dialogue System" *Proceedings of the ACL Interactive Poster and Demonstration Sessions, Ann Arbor*, 2005

[10] http://www.cstr.ed.ac.uk/projects/unisyn

[11] A.K. Syrdal and M. Jilka: "To Rise or To Fall: That is the Question" *Acoustical Society of America - 146th Meeting, Austin, TX*, 2003

[12] A.K. Syrdal and M. Jilka: "Acceptability of Variations In Question Intonation in Natural And Synthesized American English" *J. of the Acoustic Society of America, Vol 155, No 33*, 2004