ON CUSTOMIZING PROSODY IN SPEECH SYNTHESIS: NAMES AND ADDRESSES AS A CASE IN POINT

Kim E. A. Silverman

Artificial Intelligence Laboratory
NYNEX Science and Technology, Inc.
500 Westchester Avenue
White Plains, New York 10604

1. ABSTRACT

This work assesses the contribution of domain-specific prosodic modelling to synthetic speech quality in a name-and-address information service. A prosodic processor analyzes the textual structure of labelled input strings, and inserts markers which specify the intended prosody for the DECtalk text-to-speech synthesizer. These markers impose discourse-level prosodic organization, annotate the information structure, and adapt the speaking rate to listeners in real time. In a quantitative comparison of this domain-specific modelling with the default rules in DECtalk, the domain-specific prosody was found to reduce the transcription error rate from 14.6% to 6.4%, reduce the number of repeats requested by listeners from 2.6 to 1.1, and to sound significantly easier to understand and more natural. This result demonstrates the importance of prosodic modelling in synthesis, and implies an even more important role for prosody in more complicated domains and discourse structures.

2. INTRODUCTION

Text-to-speech synthesis could profitably be used to automate or create many information services, if only it were of better quality. Unfortunately it remains too unnatural and machine-like for all but the simplest and shortest texts. It has been described as sounding monotonous, boring, mechanical, harsh, disdainful, peremptory, fuzzy, muffled, choppy, and unclear. Synthesized isolated words are relatively easy to recognize, but when these are strung together into longer passages of connected speech (phrases or sentences) then it is much more difficult to follow the meaning: the task is unpleasant and the effort is fatiguing [1].

This less-than-ideal quality seems paradoxical, because published evaluations of synthetic speech yield intelligibility scores that are very close to natural speech. For example, Greene, Logan and Pisoni [2] found the best synthetic speech could be transcribed with 96% accuracy; the several studies that have used human speech tokens typically report intelligibility scores of 96% to 99% for natural speech. (For a review see [1]).

However, segmental intelligibility does not always predict comprehension. A series of experiments [3] compared two high-end commercially-available text-to-speech systems on application-like material such as news items, medical benefits information, and names and addresses. The result was that the one with the significantly *higher* segmental intelligibility had the *lower* comprehension scores.

Although there may be several possible reasons for segmental intelligibility failing to predict comprehension, the current work focuses on the single most likely cause: synthesis of prosody. Prosody is the organization imposed onto a string of words when they are uttered as connected speech. It includes pitch, duration, pauses, tempo, rhythm, and every known aspect of articulation. When the prosody is incorrect then at best the speech will be difficult or impossible to understand [4], at worst listeners will be *mis*understand it with being aware that they have done so.

Arguments for the importance of prosody in language abound in the literature. However, the cited examples of prosodic resolution of ambiguity usually are either anecdotal citations or are illustrated by small sets of carefully-constructed cited sentences. It is not clear how important prosody is in more normal everyday texts. This brings us to the first question addressed in the current study: how much will prosody contribute to perception of synthetic speech for non-contrived, real-world textual material?

2.1. Current Approaches to Prosody in Speech Synthesis

Text-to-speech systems are typically designed to cope with "unrestricted text" [5]. Each sentence in the input text is analyzed independently, and the prosody that is applied is a trade-off to avoid one the one hand not sounding too monotonous, and on the other hand implementing the prosodic features so saliently that egregious errors occur when the wrong prosodic features are applied. The approach taken in these systems to generating the prosody has been to derive it from an impoverished syntactic analysis of the text to be spoken. Usually content words receive pitch-related prominence, function words do not. Small prosodic

boundaries, marked with pitch falls and some lengthening of the syllables on the left, are inserted wherever there is a content word on the left and a function word on the right. Larger boundaries are placed at punctuation marks, accompanied by a short pause and preceded by either a falling-then-rising pitch shape to cue nonfinality in the case of a comma, or finality in the case of a period. Declination of pitch is imposed over the duration of each sentence.

There are several ways in which deviations from the above principles can be implemented to add variety and interest to an intonation contour. For example the declination may be partially reset at commas within a sentence. Or the extent of prominence-lending pitch excursions on content words may be varied according to their lexical class (higher pitch peaks on nouns or adjectives, lower on verbs) or their position in the phrase (alternating higher and lower peaks). These variations may be based on stochastically trained models.

One problem with the above approach is that prosody is not a lexical property of English words — English is not a tone language. Neither is prosody completely predictable from English syntax — prosody is not a redundant encoding of already-inferable information.

Rather, prosody annotates the information structure of the accompanying text string. It depends on the prior mutual knowledge of the speaker and listener, and on the role a particular utterance takes within its particular discourse. It marks which concepts are considered by the speaker to be new in the dialogue, which ones are topics, and which ones are comments. It encodes the speaker's expectations about how the current utterance relates to that the listener's current knowledge, it indicates focussed versus background information. This realm of information is very difficult to derive in an unrestricted text-to-speech system, and it is correspondingly difficult to generate correct discourse-relevant prosody. This is a primary reason why long passages of synthetic speech sound so unnatural.

2.2. Application-specific discourse constraints on prosody

There are many different applications for synthetic speech, but what they tend to share in common is that usually within each application (i) the text is *not* unrestricted, but rather is a constrained topic and a limited subset of the language, and (ii) the speech is spoken within a known discourse context. Therefore within the constraints of a particular application it is possible to make assumptions about the type of text structures to expect, the reasons the text is being spoken, and the expectations of the listener. These are just the types of information that are necessary to constraint the prosody. This brings us to the second aim of the current research: is it possible to create application-specific rules to improve the prosody in a real text-to-speech synthesis application?

Prior work has shown that discourse characteristics of simulated applications can be used to constrain prosody. Young and Fallside [6] built a system that enabled remote access to status information about East Anglia's water supply system. This system answered queries by generating text around numerical data and then synthesizing the resulting sentences. The desired prosody was generated along with the text, rather than being left to the default rules of an unrestricted text-to-speech system. Silverman developed paragraph-level rules to vary pitch range and place accents based on a model of recently-activated concepts. Hirschberg and Pierrehumbert [7] generated the prosody in synthetic speech according to a block structure model of discourse in an automated tutor for the `vi` text editor. Davis [8] built a system that generated travel directions within the Boston metropolitan area. In one version of the system, elements of the discourse structure (such as given-versus-new, repetition, and grouping of sentences into larger units) were used to manipulate accent placement, boundary placement, and pitch range.

Each of these pieces of research consists of a carefully-elaborated set of rules to improve synthetic speech quality. However the evidence that the speech did indeed sound better was more intuitive than based on formal perceptual assessments. Yet systematic and controlled evaluation is crucial in order to test whether hypothesized rules are correct, and whether they have a measurable effect on how the speech is perceived.

The current work builds on the progress made in the above systems by evaluating prosodic modelling in the context of an existing information-provision service.

3. PROSODY FOR A NAME AND ADDRESS INFORMATION RETRIEVAL SERVICE

The text domain for the current work is synthesis of names and addresses. The associated pronunciation rules and text processing are well understood, and there are many applications that require this type of information. At the same time this represents a particularly stringent test for the contribution of prosody to synthesis quality because names and addresses have such a simple linear structure. There is little structural ambiguity, no center-embedding, no relative clauses. There are no indirect speech acts. There are no digressions. Utterances are usually very short. In general, names and addresses contain few of the features common in cited examples of the centrality of prosody in spoken language. This class of text seems to offer little opportunity for prosody to aid perception.

On the other hand, if prosody can be shown to influence synthetic speech quality even on such simple material as names and addresses, then it is all the more likely to be important in spoken language systems where the structure of the material is more complex and the discourse is richer.

3.1. The application dialogue

This work took place within the context of a field trial of speech synthesis to automate NYNEX's reverse-directory service [9]. Callers are real users of the information service. They know the nature of the information provision service, before they call. They have 10-digit telephone numbers, for which they want the associated listing information. At random, their call may arrive at the automated position. The dialogue with the automated system consists of two phases: information gathering and information provision. The information-gathering phase used standard Voice Response Unit technology: they hear recorded prompts and answer questions by pressing DTMF keys on their telephones. This phases establishes features of the discourse that are important for generating the prosody: callers are aware of the topic and purpose of the discourse and the information they will be asked to supply by the interlocutor (in this case the automated voice). It also establishes that the interlocutor can and will use the telephone numbers as a key to indicate how the to-be-spoken information (the listings) relates to what the caller already knows (thus "*555 1234 is listed to Kim Silverman, 555 2345 is listed to Sara Basson*").

The second phase is information provision: the listing information for each telephone number is spoken by a speech synthesizer. Specifically, the number and its associated name and town are embedded in carrier phrases, as in:
<number> is listed to <name> in <town>
The resultant sentence is spoken by the synthesizer, after which a recorded human voice offers to repeat the listing, spell the name, or continue to the next listing.

These features may seem too obvious to be worthy of comment, but they very much constrain likely interpretations of what is to be spoken, and similarly define what the appropriate prosody should be in order for the to-be-synthesized information to be spoken in a compliant way.

3.2. Rules for Prosody in Names and Addresses

In the field trial, text fields from NYNEX's Customer Name and Address database (approximately 20 million entries) are sent to a text processor [10] which identifies and labels logical fields, corrects many errors, and expands abbreviations. For the current research, a further processor was written which takes the cleaned-up text which is output from that text processor, analyzes its information structure, and inserts prosodic markers into it before passing it on to a speech synthesizer. The prosodic markers control such things as accent type, accent location, overall pitch range, boundary tones, pause durations, and speaking rate. These are recognized by the synthesizer and will override that synthesizer's own inbuilt prosody rules.

The prosodic choices were based on analyses of 371 interactions between real operators and customers. The operators use a careful, clear, deliberately-helpful style when saying this information. The principles that underlie their choice of prosody, however, are general and apply to all of language. The tunes they use appear to be instances of tunes in the repertoire shared by all native speakers, their use of pitch range is consistent with observational descriptions in the Ethnomethodology literature, their pauses are neither unrepresentatively long nor rushed. What makes their prosody different from normal everyday speech is merely which tunes and categories they select from the repertoire, rather than the contents of the repertoire itself. This reflects the demand characteristics of the discourse.

The synthesizer which was chosen for this prosodic preprocessor was DECtalk, within the DECvoice platform. This synthesizer has a reputation for very high segmental intelligibility [2]. It is widely used in applications and research laboratories, and has an international reputation.

There are three categories of processing performed by the prosodic rules: (i) discourse-level shaping of the overall prosody; (ii) field-specific accent and boundary placement, and (iii) interactive adaptation of the speaking rate.

(i) Discourse-level shaping of the prosody within a turn. That turn might be one short sentence, as in **914 555 2145 shows no listing**, or several sentences long, as in **The number 914 555 2609 is an auxiliary line. The main number is 914 555 2000. That number is handled by US Communications of Westchester doing business as Southern New York Holdings Incorporated in White Plains NY 10604.** The general principle here is that prosodic organization can span multiple intonational phrases, and therefore multiple sentences. These turns are all prosodically grouped together by systematic variation of the overall pitch range, lowering the final endpoint, deaccenting items in compounds (e.g. "*auxiliary* line"), and placing accents correctly to indicate backward references (e.g. "*That* number..."). The phone number which is being echoed back to the listener, which the listener only keyed in a few seconds prior, is spoken rather quickly (the 914 555 2145, in this example). The one which is new is spoken more slowly, with larger prosodic boundaries after the area code and local exchange, and an extra boundary between the eighth and ninth digits. This is the way native speakers say this type of information when it is new and important in the discourse.

Another characteristic of this level of prosodic control is the type and duration of pauses within and between some of the sentences. Some pauses are inserted within intonational phrases, immediately prior to information-bearing words. These pauses are NOT preceded by boundary-related pitch tones, and only by a small amount of lengthening of the preceding material. They serve to alert the listener that something important is about to be spoken, thereby focussing the listener's attention. In the TOBI transcription system, these would be transcribed as a 2 or 2p boundary. Example locations of these pauses include: **"The main number is... 914 555 2000."** and **"In... White Plains, NY 10604."**

The duration of the sentence-final pause between names and their associated addresses is varied according to the length and complexity of the name. This allows listeners more time to finish processing the acoustic signal for the name (to perform any necessary backtracking, ambiguity resolution, or lexical access) before their auditory buffer is overwritten by the address.

(ii) Signalling the internal structure of labelled fields. The most complicated and extensive set of rules is for name fields. Rules for this field first of all identify word strings which are inferable markers of information structure, rather than being information-bearing in themselves, such as "... doing business as...". The relative pitch range is reduced, the relative speaking rate is increased, and the stress is lowered. These features jointly signal to the listener the role that these words play. In addition, the reduced range allows the synthesizer to use its normal and boosted range to mark the start of information-bearing units on either side of these markers. These units themselves are either residential or business names, which are then analyzed for a number of structural features. Prefixed titles (Mr, Dr, etc.) are cliticized (assigned less salience so that they prosodically merge with the next word), unless they are head words in their own right (e.g. "Misses Incorporated"). Accentable suffixes (incorporated, the second, etc.) are separated from their preceding head and placed in an intermediate-level phrase of their own. After these are stripped off, the right hand edge of the head itself is searched for suffixes that indicate a complex nominal. If one of these is found is has its pitch accent removed, to yield for example Building Company, Plumbing Supply, Health Services, and Savings Bank. However if the preceding word is a function word then they are NOT deaccented, to allow for constructs such as "John's Hardware and Supply", or "The Limited". The rest of the head is then searched for a prefix on the right, in the form of "<word> and <word>". If found, then this is put into its own intermediate phrase, which separates it from the following material for the listener. This causes constructs like "A and P Tea Company" to NOT sound like "A, and P T Company" (prosodically analogous to "A, and P T Barnum").

Within a head, words are prosodically separated from each other very slightly, to make the word boundaries clearer. The pitch contour at these separations is chosen to signal to the listener that although slight disjuncture is present, these words cohere together as a larger unit.

Similar principles are applied within the other address fields. In address fields, for example, a longer address starts with a higher pitch than a shorter one, deaccenting is performed to distinguish "Johnson Avenue" from "Johnson Street", ambiguities like "120 3rd Street" versus "100 23rd Street" versus "123rd Street" are detected and resolved with boundaries and pauses, and so on. In city fields, items like "Warren Air Force Base" have the accents removed from the right hand two words.

An important component of signalling the internal structure of fields is to mark their boundaries. Rules concerning inter-field boundaries prevent listings like "Sylvia Rose in Baume Forest" from being misheard as "Sylvia Rosenbaum Forest".

(iii) Adapting the speaking rate. Speaking rate is a powerful contributor to synthesizer intelligibility: it is possible to understand even an extremely poor synthesizer if it speaks slowly enough. But the slower it speaks, the more pathological it sounds. Moreover as listeners become more familiar with a synthesizer, they understand it better and become less tolerant of unnecessarily-slow speech. Consequently it is unclear what the appropriate speaking rate should be for a particular synthesizer, since this depends on the characteristics of both the synthesizer and the application.

To address this problem, a module modifies the speaking rate from listing to listing on the basis of whether customers request repeats. Briefly, repeats of listings are presented faster than the first presentation, because listeners typically ask for a repeat in order to hear only one particular part of a listing. However if listener consistently requests repeats for several consecutive listings, then the starting rate for new listings within that call is slowed down. If this happens over sufficient consecutive calls, then the default starting rate for a new call is slowed down. Similarly, if over successive listings or calls there are no repeats, then the speaking rate will be increased again. By modelling three different levels of speaking rate in this way (within-listing, within-call, and across-calls), this module attempts to distinguish between a particularly difficult listing, a particularly confused listener, and an altogether-too-fast (or too slow) synthesizer.

In addition to the above prosodic controls, there is a specific module to control the way items are spelled when listeners request spelling This works in two ways. Firstly, using the same prosodic principles and features as above, it employs variation in pitch range, boundary tones, and pause durations to define the end of the spelling of one item from the start of the next (to avoid "Terrance C McKay Sr." from being spelled "T-E-R-R-A-N-C-E-C, M-C-K-A Why Senior"), and it breaks long strings of letters into groups, so that "Silverman" is spelled "S-I-L, V-E-R, M-A-N". Secondly, it spells by analogy letters that are ambiguous over the telephone, such as "F for Frank", using context-sensitive rules to decide when to do this, so that it is not done when the letter is predictable by the listener. Thus N is spelled "N for Nancy" in a name like "Nike", but not in a name like "Chang". The choice of analogy itself also depends on the word, so that "David" is NOT spelled "D for David, A,...."

4. PRELIMINARY EVALUATION

A transcription experiment was carried out to evaluate the impact of the prosodic rules on the synthetic speech quality

in terms of both objective transcription accuracy and of subjective ratings.

4.1. Test material

A set of twenty-three names and addresses had been already been developed by Sara Basson (unpublished ms, 1992) for assessing the accuracy with which listeners can transcribe such material. This set had been constructed to represent the variation in internal structure and length that occurred in NYNEX's database. Although it did contain some material that would be ambiguous if synthesized with incorrect prosody, it was not intended to focus exclusively on prosodic variability and was developed before the prosodic processor was finished. It contained phonemic diversity;, a variety of personal names, cities and states; short and long name fields, and digit strings. There were roughly equal proportions of easy, moderate, and difficult listings, as measured by how well listeners could transcribe the material when spoken by a human. Henceforth each of these names and addresses shall be referred to as *items*.

4.2. Procedure

The 23 items were divided into two sets. Listeners were all native speakers of English with no known hearing loss, and all employees of NYNEX Science and Technology. On the basis of our previous experience with synthetic speech perception experiments, we expect these listeners will perform better on the transcription task than general members of the public. Thus the results of this transcription test represent a "best case" in terms of how well we can expect real users to understand the utterances.

Listeners called the computer over the public telephone network from their office telephones: their task was to transcribe each of the 23 items. Each listener heard and transcribed the items in two blocks: one of the sets of items spoken by DECtalk's default prosody rules, and the other spoken with application-specific prosody. The design was counter-balanced with roughly half of the listeners hearing each version in the first block, and roughly half hearing each item set in the first block. For each item, listeners could request as many repeats as they wanted in order to transcribe the material as accurately as they felt was reasonably possible. Listeners were only allowed to request spelling in two of the items, which were constructed to sound like pronounceable names and contain every letter in the alphabet.

4.3. Dependent variables

Transcription scores per item. Each word in each item could score up to 3 points. One point would be deducted if the right-hand word boundary was misplaced, one point if one phoneme was wrong, and two points of more than one phoneme was wrong.

Number of repeats requested per item. For items that were spelled, this was the number of times after the first spelling.

Perceived intelligibility. Each version of the synthesis was rated by each listener on a five-point scale labelled: "How easy was it to understand this voice?" (where 1 = "Consistently failed to understand much of the speech" and 5 = "Consistently effortless to understand").

Perceived naturalness. Each version was similarly rated, on a five-point scale labelled "How natural (i.e. like a human voice) did this voice sound? (where 1 = extremely unnatural and 5 = extremely natural).

Preferences. Since each listener heard each voice, they were asked for which voice they preferred: voice 1, voice 2, or no preference.

4.4. Results

So far results have been analyzed for 17 listeners. Summing over all transcriptions, the maximum possible transcription score for each synthesizer was 5032. The per-word error rate for items spoken with the synthesizer's default prosody was 14.6%. With the domain-specific prosody this was only 6.4%. Thus listeners could transcribe the vowels and consonants significantly more accurately even though the vowels and consonants are pronounced by exactly the same segmental rules in both cases. The only difference is the prosody.

Transcription scores do not reflect how much effort listeners expended to achieve their transcription accuracy. One measure of that effort is the number of repeats they requested. Listeners needed on average 2.6 repeats per listing for the default prosody, but only 1.1 repeats per listing with the domain-specific prosody. Interestingly, in a prior transcription test with a *human* voice saying a superset of the listings used in this experiment, listeners needed 1.2 repeats per listing (Sara Basson, personal communication).

On the "ease of understanding" scale, the default prosody scored 1.8 (standard deviation = 0.8), while domain-specific prosody scored 3.3 (standard deviation = 0.8). Thus listeners' subjective perceptions matched their objective transcription results: they were aware that the version with domain-specific prosody was easier to understand, though clearly it was not effortless.

On the "naturalness" scale, the default prosody scored 1.9 (standard deviation = 0.9) and domain-specific prosody scored 2.9 (standard deviation = 0.8). Though statistically significant, this difference is smaller than on the previous scale. Alteration of the just the pitch and duration made the

speech made the speech sound somewhat more natural, but it is still is a long way from sounding "extremely natural".

One the preference ratings, so far all of the listeners preferred the speech versions with domain-specific prosody.

5. CONCLUSION

Although this evaluation is preliminary, it suggests that even in such simple material as names and addresses domain-specific prosody can make a clear improvement to synthetic speech quality. The transcription error rate was more than halved, the number of repetitions was more than halved, the speech was rated as more natural and easier to understand, and it was preferred by listeners. This result encourages further research on methods for capitalizing on application constraints to improve prosody. The principles in the literature for customizing the prosody will generalize to other domains where the structure of the material and discourse purpose can be inferred.

The second conclusion is that at least in this domain, although domain-specific rules can improve synthetic prosody over that in domain-independent rules, the domain-specific customization can be severely limited if the synthesizer does not make the right prosodic controls available. In an ideal world, the markers that are embedded in the text would specify exactly how the text is to be spoken. In reality, however, they specify at best an approximation. This exercise is constrained by the controls made available by that synthesizer. Some manipulations that are needed for this type of customization are not available, and some of the controls that are available interact in mutually-detrimental ways. Consequently to the extent that the application-specific prosody did indeed improve synthesis quality, this is all the more supporting evidence for both the importance of generating domain-relevant prosody on the one hand, and for NOT doing it with such an improper prosodic model on the other.

The immediate next steps in this work are to more systematically evaluate the perceptual impact of the above rules, both in transcription tests and with the quantitative measures of acceptance by real users that are already being used in the field trial. In addition, we are currently developing a set of rules to customize the prosody in a spoken language system for remote financial transactions, combining text-specific rules of the type evaluated in this work, with rules that will use the discourse history to dynamically derive information about topics, discourse functions of replies, and given versus new information.

The development and evaluation of this work furthers our understanding of (i) how to use prosody to clarify names and addresses in particular, and other texts in general; (ii) prosody's importance in a real application context, rather than in laboratory-generated unrepresentative sentences; (iii) one way to incorporate user-modelling of speaking rate into speech synthesis (speakers should not ignore their listeners); and (iv) what prosodic controls a synthesizer should make available.

6. ACKNOWLEDGEMENTS

This work could not have proceeded without the context and focus of the ACNA trial in general, and in particular the efforts and insights of Dina Yashchin, Ashok Kalyanswamy, Sara Basson, John Pitrelli, and Judy Spitz. Shortcomings of course remain my own responsibility.

REFERENCES

1. Silverman, K.E.A. *The Structure and Processing of Fundamental Frequency Contours.* Ph.D. Dissertation, Cambridge University, 1987.

2. Greene, B.G.; Logan, J.S. and Pisoni, D.B. "Perception of synthetic speech produced automatically by rule: Intelligibility of eight text-to-speech systems", *Behavior Research Methods, Instruments, and Computers*, Vol. **18**, 1986, pp 100-107

3. Silverman, K.E.A., Basson, S. and Levas, S. Evaluating Synthesizer Performance: Is Segmental Intelligibility Enough? Proc. ICSLP-90, Vol. **1**, 1990.

4. Huggins, A.W.F. "Speech Timing and Intelligibility. In J. Requin (Ed): *Attention and Performance VII.* Erlbaum, Hillsdale. 1978.

5. Allen, J, Hunnicutt, M.S., Klatt, D., Armstrong, R.C. and Pisoni, D.B. *From Text to Speech: The MITalk System.* Cambridge University Press, Cambridge, 1987

6. Young, S.J. and Fallside, F. "Synthesis by rule of prosodic features in word concatenation synthesis", *Int. J. Man-Machine Studies*, Vol. **12**, 1980, pp 241-258.

7. Hirschberg, J. and Pierrehumbert, J.B. "The Intonational Structuring of Discourse", *Proc. 24th ACL Meeting*, 1986, pp 136-144.

8. Davis, J.R. "Generating intonational support for discourse", *J. Acoust. Soc. Am. Suppl. 1*, Vol. **82**, 1987, p S17.

9. Yashchin, D, Basson, S., Kalyanswamy, A., Silverman, K.E.A. "Results from automating a name and address service with speech synthesis". *Proc AVIOS-92*, 1992.

10. Kalyanswamy, A. and Silverman, K.E.A. "Processing information in preparation for text-to-speech synthesis". *Proc AVIOS-92*, 1992.