Improving the Assignment of Lexical Stress in Text-to-Speech Systems

Joanne Arciuli[1,2] and James Thompson[2]

[1]Department of Psychology
Charles Sturt University,
Australia
jarciuli@csu.edu.au

[2]Centre for Research in Complex Systems (CRiCS)
Charles Sturt University,
Australia
jthompson@csu.edu.au

Abstract

Despite intensive research effort, humans still produce speech with greater accuracy than machines. In addition to clear gaps in accuracy, machine speech (also known as speech synthesis or text-to-speech) is renowned for its lack of 'naturalness' – especially in terms of prosody. Here we demonstrate how recent research into human language processing may improve the accuracy and naturalness of machine speech, particularly with regard to stress assignment in individual words.

1. Introduction

Lexical stress refers to the opposition of stressed and unstressed syllables within polysyllabic words (e.g., the word 'zebra' has more emphasis on the first syllable than the second syllable). Lexical stress is an important feature of English and it has been found that incorrect stress assignment significantly affects intelligibility (Slowiaczek, 1990).

For researchers examining the conversion of print to speech in humans - reading aloud - there has been a strong focus on the processing of monosyllabic words and, as a result, stress assignment has been largely neglected. However, there have been recent efforts to shed light on stress assignment during reading (e.g., Arciuli & Cupples, 2006; in press). This recent research demonstrates that there is a rich source of probabilistic cues to stress in the orthography of English. Large-scale analyses of English disyllables in the CELEX database have revealed reliable cues in words' beginnings (e.g., 'tu-', 'sta-', 'li-' are typically seen in words with first-syllable stress while 'i-', 'de-', 'be-' are mostly found in words with second-syllable stress) and endings (e.g., '-ip', '-ock', '-us' are typically seen in words with first-syllable stress while '-act', '-ibe', '-oin' are often found in words with second-syllable stress). These cues extend well beyond the limited set of known prefixes and affixes. Moreover, behavioural tests using nonsense words with biasing beginnings/endings demonstrate that participants are sensitive to these cues when assigning stress. It is noteworthy that much of this research has focussed on nouns and verbs (the largest grammatical categories in English).

On the technology side, researchers working on the development of text-to-speech systems (speech synthesis – or machine speech) have not had the luxury of ignoring stress assignment. Text-to-speech engines (TTS) deal with polysyllabic words constantly and must integrate procedures which assign stress within individual words. One popular open source text-to-speech system utilises letter-to-sound rules (LTS) to generate a combined phonetic and prosodic code from an orthographic representation. In their current form, the LTS rules predict stress for each syllable (not the word as a whole), taking into account: the syllable position in the word, vowel length, vowel height, and the number of syllables from the end of the word. The LTS rules were developed using a 'Classification And Regression Tree' (CART) algorithm, and are implemented in three major TTS systems (Festival, Flite and FreeTTS – Walker, Lamere & Kwok, 2002; http://freetts.sourceforge.net/docs/index.php).While there are a variety of commercial products available our study focuses on this system only.

In the current study we sought to determine whether the probabilistic cues discovered during research on human language processing could assist in

Proceedings of the 11th Australian International Conference on Speech Science & Technology, ed. Paul Warren & Catherine I. Watson. ISBN 0 9581946 2 9
University of Auckland, New Zealand. December 6-8, 2006.

Accepted after abstract only review

improving stress assignment in TTS. Given the complexity of LTS rules that have been devised to account for stress assignment in English it is of particular interest to ascertain whether something as straightforward as attending to the (non-morphological) orthography at the beginning and ending results in better accuracy of output. Our approach was as follows: 1) Identify the limitations of TTS with regard to stress assignment, 2) Identify the potential advantage of using words' beginnings and endings to assign stress by building a neural network. For both parts of this research we focused on a set of 120 real English disyllables - used in a previously published study conducted by Arciuli and Cupples (2006). The set consists of 60 words with first-syllable stress and 60 with second-syllable stress (120 words in total). Half the items are low in frequency and half are high in frequency and there are equal numbers of nouns and verbs. The set includes words such as 'courage', 'garment', 'arrive', and 'punish'. The entire set of words is included in Appendix A.

2. Experiment One - Identifying the Limitations of TTS

Accurate and 'natural' stress assignment is a particular problem for TTS. Indeed, when we ran our stimuli through the TTS system and looked at the stress assignment values separately from the phonemic output we found that it correctly assigned stress for only 64% of words. The errors were of three kinds. A total of 12% of the errors involved assigning stress to the wrong syllable (e.g., suggesting 'proDUCT' with second-syllable stress instead of 'PROduct' with first-syllable stress). A total of 32% of errors involved a failure to assign any stress (e.g., 'defence' without second-syllable stress). The remaining 56% involved assigning stress to both syllables (i.e., even stress) where there should have been stronger stress on just one of those syllables (e.g., 'SOLVENT'). This data was obtained by looking at text output concerning the acoustic representations – where stress and phonemic information are coded separately.

Of course, it is important to ascertain how participants' perceive the acoustic output. It may be that 'close enough is good enough' the majority of the time and that participants do not actually perceive incorrect stress assignment in the TTS acoustic output. However, it is also possible that participants do perceive these incorrect stress patterns and/or report the stress patterns as being 'unnatural'.

2.1. Methods

Eighteen monolingual speakers of English speakers from the undergraduate Psychology course participated in exchange for monetary compensation. Participants reported normal hearing.

We obtained TTS acoustic output (male voice with US accent – there was no Australian accent available) for the 120 words (stimuli set detailed above) and created a random ordering. Each word was preceded by a number. There was a one second delay between the number and the word and five seconds delay between each trial.

Participants were tested in one group session - they heard the same random order and were given the same written response sheet. Their task was to listen to each item and identify where the stress was in the word ("beginning", "end" or "can't decide") and to rate the naturalness of the overall pronunciation on a five-point Likert scale (1=Very Unnatural, 2=Unnatural, 3=Neutral, 4=Natural, 5=Very Natural).

Participants were given three practice items which they responded to the presence of the experimenter. During this practice phase, the experimenter explained the nature of the task and answered any questions.

2.2. Results

We examined each item separately and determined: 1) the percentage of participants that reported stress placement in the correct position, and 2) the naturalness rating.

Analysis of the data showed that, on average, only 45.32% of participants reported stress placement in line with the Received Pronunciation for each word. This figure lends support to the suggestion that the TTS system does not assign stress adequately.

Further analysis revealed an average naturalness rating of 2.89. This indicates that participants were somewhat undecided regarding the naturalness of pronunciation but erred towards rating items as sounding unnatural.

We returned to our text output from the TTS system and coded the participant data depending on whether the textual output showed incorrect stress placement or not. We then ran a one-way ANOVA using percentage of participants that perceived stress in line with Received Pronunciation as the dependent variable. Results indicated a significant difference in participants' perception of words classified as having incorrect stress placement in the textual output vs. words that showed correct stress placement in the textual output $F(1,118) = 24.05$, $p < .0005$). This indicates that words that were listed as having incorrect stress placement in the textual output tended to be perceived as having incorrect stress placement. This is depicted in the Figure 1.

A second one-way ANOVA using naturalness ratings as the dependent variable revealed a significant difference between words that were listed in the textual output as having incorrect stress vs. words with correct stress $F(1,118) = 15.27$, p 0005). Words that were listed as having incorrect stress placement in the textual output tended to be rated as sounding less natural. This is depicted in the Figure 2.

Proceedings of the 11th Australian International Conference on Speech Science & Technology, ed. Paul Warren & Catherine I. Watson. ISBN 0 9581946 2 9
University of Auckland, New Zealand. December 6-8, 2006.

Accepted after abstract only review

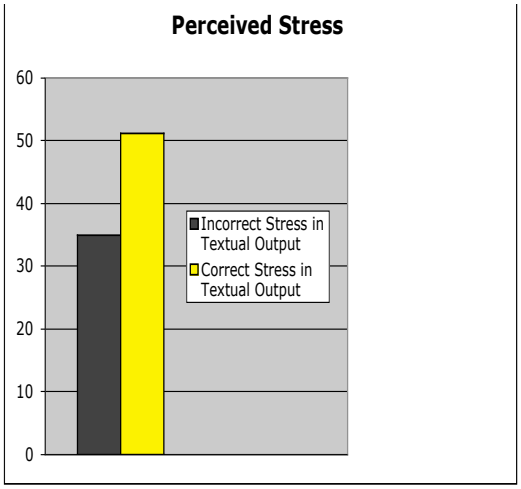

Figure 1. Percentage of participants that perceived stress in line with Received Pronunciation

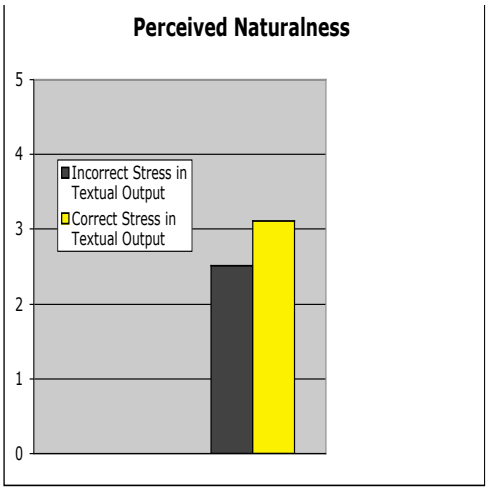

Figure 2. Naturalness Ratings (1 = Very Unnatural; 5 = Very Natural)

3. Experiment Two - Neural Network

Based on the earlier findings of Arciuli and Cupples (2006; in press) we created a neural network (NN) that was designed to assign stress patterns on the basis of words' beginnings and endings.

3.1. Methods

The words for the training data were extracted from the CELEX database (Baayen, Pipenbrock & Gulikers, 1995). Words were selected if they met the following selection criteria: identified by CELEX as having two syllables, listed as a single wordform (e.g. "complete" and not "tag line"), and lastly, classified by CELEX as either a noun of verb (identical words fitting both categories were included twice - e.g "hammer" was included twice as a verb and a noun). We focused on nouns and verbs as earlier work by Arciuli and Cupples had focused on these categories.

This selection criteria resulted in 10,771 words being selected. Words were then separated into their beginnings and endings by an algorithm. Beginnings were defined as all the letters up to and including the first vowel or vowel cluster (e.g. in "saucy" the beginning would be "sau", and in "listen" it would be "li"). Endings were identified by working from the end of the word to the beginning and segmenting all the letters up to and including the first vowel or vowel cluster (e.g. in "saucy" the end would be "y", in "listen" it would be "en", etc). These definitions are from the earlier work by Arciuli and Cupples.

Finally, to simplify the structure of our NN, we excluded words where the number of letters in either the beginning or ending exceeded four, reducing the total number of words to 10,134.

The NN used for this experiment was a simple three layer feed forward network with 216 inputs, 28 nodes in the hidden layer and 4 output nodes.

The input layer was made up of 8 rows of 27 nodes, one row of 4 for beginnings and another row of 4 for the endings. In each row there was a node for each possible letter of the alphabet including a 'null' node representing the lack of a letter in that position. The rows were filled from the first row of four to the last for each word, e.g. for the word "saucy" the input would have looked like SAU0Y000, and for "listen" LI00EN00 (see Figure 3).

The output layer was made up of one row of 4 nodes, with the first node representing first syllable stress, the second indicating second syllable stress (the third indicated noun status and the fourth indicated verb status – but this data is being used for another project).

The extracted word beginnings and endings were converted into input patterns to train the neural network. The network was trained for 1000 epochs, where at each epoch the neural network was trained using all 10,134 words from the list, presented in random order. The neural network was then tested against the word list in Appendix A in a random order.

A word's stress was assigned by allowing a threshold for difference between the first output node (N_1) and output second node (N_2) of the output layer, such that if $N_1 > (N_2 + 0.05)$ the stress would be on the first syllable, if $N_2 > (N_1 + 0.05)$ the stress would be on the second syllable, and if $\sqrt{(N_1 - N_2)^2} < 0.05$ then no stress would be assigned.

Proceedings of the 11th Australian International Conference on Speech Science & Technology, ed. Paul Warren & Catherine I. Watson. ISBN 0 9581946 2 9
University of Auckland, New Zealand. December 6-8, 2006.

Accepted after abstract only review

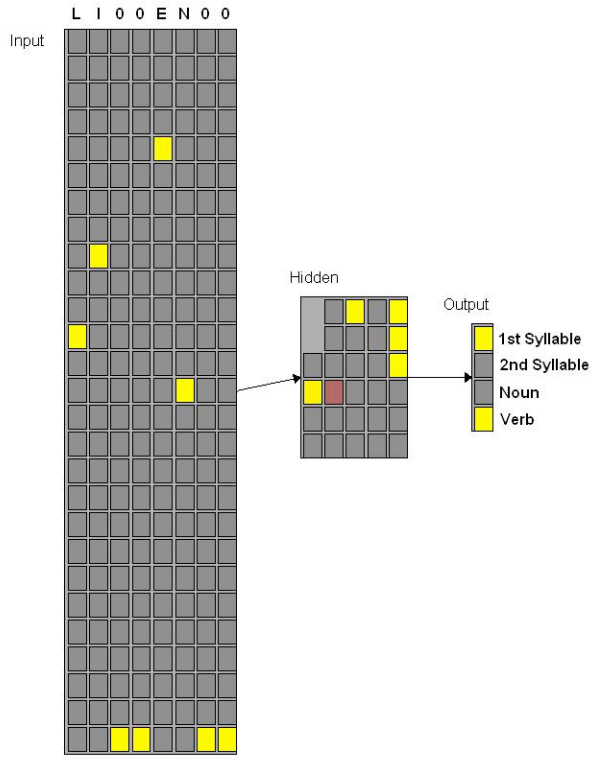

Figure 3. Structure of NN

3.2. Results

The network handled stress assignment far more accurately than the TTS system. Overall, the NN correctly assigned stress 86% of the time. The errors were of two kinds. A total of 79% of the errors involved incorrect stress assignment (e.g., suggesting 'effORT' with second-syllable stress instead of 'Effort' with first-syllable stress). The remaining 21% of errors involved a failure to assign any stress (e.g., 'aspect' without first-syllable stress).

4. Discussion

The research presented here clearly demonstrates: 1) The limitations of TTS with regard to stress assignment, 2) The advantage of using words' beginnings and endings to assign stress (i.e., non-morphological orthography).

Using a set of 120 stimuli we showed that a popular and well-respected open-source TTS system achieved 64% accuracy on stress assignment. Behavioural testing confirmed that listeners had problems perceiving stress patterns – on average, less than half the participants perceived stress patterns that were in line with Received Pronunciation. Moreover, participants erred towards rating overall pronunciation as unnatural. Our neural network used words' beginnings and endings (derived from non-morphological processes) to predict stress patterns and achieved an impressive accuracy rate of 86%. Admittedly, our NN was only predicting stress patterns, whereas the TTS systems must engage in the far more challenging exercise of producing acoustic output. However, when one considers accuracy of stress assignment this research demonstrates that complex 'rules' that include assessment of vowel length and vowel height etc may not be as accurate as previously thought.

This research suggests that TTS systems might be improved if they were to incorporate a simple non-morphological strategy regarding stress assignment – based on probabilistic cues present in the spelling patterns of words' beginnings and endings. Certainly, Arciuli and Cupples (2006; in press) have shown that there is a rich source of cues to stress in the spelling patterns of English disyllables and that participants are sensitive to these cues. Efficient use of these cues may help to explain the ease and accuracy of stress assignment in human speech production but can also be used to enhance machine speech.

We are now expanding this research to include trisyllabic words and words from all grammatical categories. We are also currently investigating the contribution made by beginnings vs endings. Preliminary results suggest that endings may contribute more information on stress assignment. Ultimately, our aim is to incorporate a probabilistic algorithm for stress assignment (based on words' beginnings and endings – and possibly weighted more heavily in favour of endings) into this open-source TTS system to ascertain whether this results in increased accuracy and naturalness.

Proceedings of the 11th Australian International Conference on Speech Science & Technology, ed. Paul Warren & Catherine I. Watson. ISBN 0 9581946 2 9
University of Auckland, New Zealand. December 6-8, 2006.

Accepted after abstract only review

5. References

Arciuli, J., & Cupples, L. (2006). The processing of lexical stress during visual word recognition: Typicality effects and orthographic correlates. *Quarterly Journal of Experimental Psychology, 59:5,* 920-948.

Arciuli, J., & Cupples, L. (in press). Would you rather 'embert a cudsert' or 'cudsert an embert'? How spelling patterns at the beginning of English disyllables can cue grammatical category. To appear in Schalley, A. & Khlentzos, D. (Eds.): *Mental States: Language and Cognitive Structure.* Accepted 2005.

Baayen, R.H., Pipenbrock, R. & Gulikers, L. (1995). *The CELEX Lexical Database* (CD-ROM). Linguistic Data Consortium, University of Pennsylvania, Philadelphia, PA.

FreeTTS Homepage

http://freetts.sourceforge.net/docs/index.php,

(Last visited: 20/10/06)

Slowiaczek, L. (1990). Effects of lexical stress in auditory word recognition. *Language and Speech, 33,* 47-68.

Walker, W., Lamere, P., & Kwok, P. (2002). FreeTTS - a performance case study. *SMLI Technical Report* 2002, 114.

6. Appendix

Experimental Stimuli

product	dagger	publish
effort	axle	carry
symbol	patron	listen
tension	leisure	settle
welfare	haven	vary
volume	cadet	marry
danger	champagne	threaten
absence	restraint	manage
future	cigar	furnish
courage	applause	suffer
region	terrain	comply
justice	platoon	excel
knowledge	sardine	impair
crisis	lapel	despise
aspect	ravine	suppress
relief	divan	amend
event	morale	invent
belief	saloon	confide
defence	domain	vacate
technique	quartet	distract
affair	afford	evolve
extent	enjoy	provoke
degree	reflect	propel
prestige	arrive	adore
advice	assert	deprive
success	oppose	nourish
response	prevent	hover
expense	emerge	injure
device	relax	punish
percent	receive	cancel
solvent	acquire	perish
kernel	forget	glisten
dungeon	allow	conquer
friction	begin	linger
stature	refer	hasten
garment	differ	grovel
treaty	gather	launder
python	follow	dazzle
demon	govern	sever
wizard	enter	shorten

Proceedings of the 11th Australian International Conference on Speech Science & Technology, ed. Paul Warren & Catherine I. Watson. ISBN 0 9581946 2 9
University of Auckland, New Zealand. December 6-8, 2006.

Accepted after abstract only review