ARCHIVES OF ACOUSTICS
32, 1, 101–109 (2007)

USING CASUAL SPEECH PHONOLOGY IN SYNTHETIC SPEECH

Linda SHOCKEY

University of Reading
Laboratory of Acoustics and Speech Communication
e-mail: linda@ias.et.tu-dresden.de

(*received October 16, 2006; accepted December 16, 2006*)

Alphabetic writing is a mixed blessing for speech science. Most scientists working in speech synthesis and speech recognition assume unconsciously that spoken language is like written language, i.e. it is composed of a string of items (letters/phonemes) which should be realised in all but substandard writing/speech. My research shows that there are very many shortcuts taken by speakers of English on a regular basis in normal (not sloppy or casual) speech. These are not included in speech synthesis packages, but if they were, the output would be closer to the real thing and, I contend, would be considerably easier to understand.

Introduction

In this paper, I will consider the importance of spontaneous speech forms in synthesis, suggesting that speech synthesis should be informed by our knowledge of human speech perception,

I. Speech Synthesis

1. Perceptual evidence of need for improvement

Evidence that synthetic speech is not perceptually equivalent to natural speech is not hard to come by. Much literature is devoted to the intelligibility of phones and sequences of phones in synthetic speech, particularly with respect to tests such as the SAM Standard Segmental Test (JEKOSCH, [8]; POLS *et al*. [16]), the Diagnostic Rhyme Test (VOIERS, SHARPLEY, HEHMSOTH, [21]) and the Modified Rhyme Test (HOUSE, WILLIAMS, HECKER, KRYTER, [5]). Other tests such as the Mean Opinion Score are used to test overall acceptability (KRAFT *et al*. [9]). Experiments by PISONI, [15] on rule-based synthetic speech show that responses to sentences to be judged True or False

are slower for synthetic speech than for natural speech and that detection accuracy similarly falls in a word spotting test.

Interpretation of the results of these tests is not categorical because synthetic speech is judged by its performance in a particular application, and not all applications require near-perfect performance. A text-to-speech system for the blind, for example, can be acceptable even though imperfect because it is significantly better than nothing. Research shows that human observers can learn strategies for understanding synthetic speech which factor in its differences from human output and that this understanding improves with practice (PISONI, [15], p. 548).

While no doubt taking this inbuilt flexibility into account, HAWKINS *et al*. [4] notes that even when individual words are relatively intelligible, connected synthetic speech puts a heavy demand on the human perceptual system. This increase on cognitive load can be attributed to "a range of poorly-modelled phenomena" such as unsatisfactory intonation contours (cf. LAURES and WEISMER, [10] in which flattening of intonation contours in natural speech reduces its intelligibility), poor stress assignment, lack of variation in rate and mismatch between prosody and information structure. All of these are suprasegmental, i.e. they have to do with the acceptability of larger patterns rather than with the accuracy of individual sounds. This suggests that speech synthesis research should 'zoom out' from its emphasis on the acoustic properties of phone-sized (and diphone-sized) segments. This is not news to those in the field, and much serious work goes into improving stress and intonation. But the idea that phonology contributes to the global acoustic profile of speech and that humans use phonology in interpreting *spans* of speech receives little attention.

2. Global perception of natural speech

Despite a tendency (perhaps by unconscious analogy with reading) to think of speech perception as sequential intrepretation of a linear string of phonological units, perception of natural speech is not based on segment-by-segment analysis. It has been known for half a century that perception of speech segments is relative to their environments (LIBERMAN *et al.*, [13], DENES, [2]). It thus makes sense for speech synthesis to be designed with the nature of the perceptual device in mind, with consideration of global as well as local features.

Research using gated[(1)] naturally-produced sentences shows that word percepts are often achieved well after the word has ended acoustically GROSJEAN, [3], for example, discovered that gated words taken from the speech stream were recognised very poorly and many monosyllabic words were not totally accepted until after their completion. LUCE, [12] agrees that many short words are not accepted until the following word is known and concludes that it is virtually impossible to recognise a word in fluent speech without first having heard the entire word as well as a portion of the next

[(1)] In gating, one truncates all but a small amount of the beginning of an utterance, then re-introduces the deleted material in small increments ('gates') until the entire utterance is heard.

word. GROSJEAN, [3] suggested that the recognition process is sequential but not always in synchrony with the acoustic-phonetic stream (though his own futher experiments showed this to be inaccurate).

BARD, SHILLCOCK, and ALTMANN, [1] presented sentences gated in words to their subjects. Although the majority of recognition outcomes (69%) yielded success in the word's first presentation with prior context only, 19% of all outcomes and 21% of all successful outcomes were late recognitions. Approximately 35% of them were identified not at the presentation of the next word, but later still. The mean number of subsequent words needed for late identification was closer to two than one (M = 1.69, SD = 1.32).

In an experiment reported in SHOCKEY, [20], I presented native English subjects with the gated sentence "So it was quite good fun, actually, on the wedding, though." I selected this sentence because the word sequence is not entirely predictable and because it showed a good deal of 'underarticulation' at the beginning.

Phonetic transcription:

[sʷ s̊ ʷ ə̊ ʷ sʷˡ kʷ waɪʔgʊ d̥ fɐ̃nætʃʊwɪɒnn̪ə̪ⁱwɛdɨŋd̪əʊ]

Here are results from two of the subjects:

1. tu
2. tut
3. uka
4. uzka
5. she's quite
6. she's quite a
7. she's quite a fun
8. she's quite a fun ac
9. she's quite good fun, ac
10. so it was quite good fun, actually ...

1. tsu
2. tsut
3. tsukɒ
4. tsuzkɒ
5. she's quite
6. she's quite a
7. she's quite a fun
9. she's quite good fun ac..
10. so it was quite good fun, actually on

It is striking that the intended word first word 'so' is detected by both perceivers at such a late point. My assumption is that it was necessary to take in this amount of information before the puzzling hissing noise at the beginning of the utterance could be interpreted: having come to an understanding of the speech rate and the temporal relationships amongst the sounds present, these two subjects could reconstitute the initial word.

I use the word 'reconstitute' advisedly, because I think speakers of English use their knowledge of phonology to replace sounds which have been removed through articulatory shortcuts, given that they are able to take in enough speech to determine general patterns. Further, I think perceivers include these shortcuts as part of their linguistic code and expect them to take place: overly carefully articulated speech is unnatural and, I suggest, harder to understand than speech containing the expected reductions.

3. Not the whole story

I do not suggest that 'phonological reconstitution by rule' is the **only** tool used in perception of reduced speech. Obviously, lexical, syntactic, pragmatic, and discourse knowledge is used, and to a greater degree as more speech is heard. Lexical knowledge rather than phonology, for example, accounts for Warren's 'phoneme replacement', in which subjects given a sentence like "The le*islature gathered in the rotunda" (where the '*' represents a cough) do not even notice that some sounds are missing. This cannot be a phonologically-related process, since it is not rule-governed.

Reconstituting 'hambag' as 'handbag' is largely attributable to phonology (due to cluster simplification and nasal assimilation), but there is supporting knowledge, such as that we do not make bags of ham and do not normally have special bags to carry ham in. If we did have these, 'hambag' would be ambiguous because of phonology, in the same way as 'hambone'. ("The hambone (=handbone) is connected to the wristbone"/"The hambone added flavour to the beans").

An example of using complementary knowlege sources to build the big picture is found in reactions to two gated sentences which I used in early experiments. These were 'The screen play didn't resemble the book' and 'The scream play was part of primal therapy'. The second word in each was pronounced 'scream', a result of nasal assimilation in the 'book' sentence. In both cases as the 'gates' opened, subjects first heard 'scream', then changed it to 'screen' *either* when the following [p] was heard (using phonological knowledge) *or* when the word 'play' was heard (probably using lexical knowlege). When the end of the second ('primal therapy') sentence arrived, some subjects changed 'screen' back to 'scream' again, in accordance with the semantics of the sentence. The advice from phonology was misleading in this case, but it obviously affected behaviour.

There seems, therefore to be evidence that phonology plays a part in speech perception in the domain of an entire sentence, in conjunction with other linguistic and real-world knowledge.

4. What are these reductions?

I include here a subset of the alternants which are part of native speaker competence in English. Several different accents are represented in the examples: Am = North American, Psmsh = Peasmarsh (Southeast England), SSB = Standard Southern British,

Cov = Coventry (Midlands, England), Ed = Edinburgh (Scotland), Nor = Norwich (East England). Most example not Am or SBS are taken from LODGE, [11].

4.1. Stress-related changes

The varieties of English I have examined depend heavily on stress as a bearer of meaning. Unstressed syllables in English tend to show reduced vowels, as is universally known. But in conversational speech, unstressed syllables undergo other kinds of reduction as well. This topic is covered at length in SHOCKEY, [20].

a. schwa absorption

I have adapted Wells' term 'schwa absorption' (1982:434) to describe cases where something else in the vicinity of a schwa takes on its syllabic property but loses the openness of a vowel, i.e. whatever sound is left has the articulatory qualities of a consonant but the syllabic qualities of a vowel. Syllabic consonants are by no means unknown for English, for example if the 't' is released nasally, the 'n' of "cotton' is syllabic, if the 't' is released laterally, the 'l' of "cattle" is syllabic. This process is extended to other cases in casual speech.

ˈfaɪn̩li	Am. "finally"	ʌnˈjʉʒɬ̩	Ed. "unusual"
ˈθaʊzn̩	Am. "thousand"	gɛʔn̩oðə	Stkpt. "get another"
ˈoʊpm̩z	Nor. "opens"	ˈlʊkŋ	Nor. "looking"
ɚɛdˈɹəʊz	'a red rose'	ɹɪˈmɛmbɹ̩ɹ̩	Psmsh. 'remember her'
ðɛɪwz	Psmsh. "they was"	wɪʧʷɚz	Am. "which was"
ʃbˈweɪsʔt	ShB. "should waste"	ˈæʧʷ	Am. "that you"
äʃtθɪŋk	Psmsh. "I should think"	ˈmækṣməm	Am. "maximum"
pʰˈlismən	Psmsh. "policemen"	pʰˈtɪkəli	Am. "particularly"

b. reduction of closure for obstruents

ˈpɛɪɸo	Stkpt. "people"	juɣŋ	SSB "you can"
ɐˈɣo	Stkpt. "I go"	bɨɣɔz	SSB "because"
pʊeˈsɛnd	Stkpt. "pretend"	ĩvæɯ̩ʔju	SSB. "in fact you"

c. tapping

ˈgɑɾˈɪn	ShB. "got in"	pʊɾəp	SSB. "put up"
ˈɛnɪbɑɾɪ	Psmsh. "anybody"	sɔɾəv	SSB. "sort of"
ˈbɛɾɪʒ	Cov. "bet his (geraniums)"	gɛɾɪn	Cov. "getting"

d. devoicing and voicing

ʋeˈleɪ v̊	Stkpt. “relieve (people)”	jɑɹts	Am. “yards (w)”
ðɪi z̊	ShB. “these” (people)	stæts	Am. “stands (n)”
ˈbæʃfəɹd̊ z̊	Psmsh. “bashfords (lived)”	ʧɛʊd̊	Ed. “child (you)”
pɾɑɖɨstənt	Ed. “protestant”	bədaɪˈθĩɣɪn	SSB. “But I think in...”
ˈgɑdə	Cov. “got a”	ˈpɑɖɨgət	Nor. “Pottergate”

4.2. Sequence-related changes

a. Cluster reduction

English is known to be a language with a potential for very heavy syllables when compared with most other languages of the world. The unmarked syllable has one initial consonant and at most one final consonant (cf McCarthy and Prince, 1994). In spontaneous speech, English moves toward the mean by reducing the number of adjacent consonants:..“a regular alternation of consonants and vowels is more natural than clusterings” (Wells, op cit: 96).

ɔ:wɪz	ShB “always”	wɐ:z	Am. “walls”
ˈweɪkes	Stkpt. “weakest”	ˈbɹɔgkɑs:əˈnjus	SSB “broadcast the news”
ˈæspɛks	SSB “aspects”	ɹɪˈspɛkfəˠ	Am. “respect for”
ˈissæɪt	Ed. “east side”	ˈdɨstɾɨks	Ed. “districts”
ɬa:ʃɪə	Cov. “last year”	ɹʌfəsˈpʰɬɛ:s	Nor. “roughest place”
fæʊnnɛm	Stkpt. “found them”	ʌʊɬˈmæn	Psmsh. “old man”
ˈtʌʊɬmi	Cov. “told me”	bæɱfəˈlaɪf	Brown, SSB “banned for life”

b. nasal relocation

tʰø̃:z	Stkpt. “turns”	dʌzə̃ʔˈwaʔ	SSB. “doesn’t want”
aθĩʔ	ShB. “I think”	kɨvĩst	Am. “convinced”
wɛw̃wɪ	ShB. “when we”	ɛĩʔ	Cov. “ain”t”
ˈfɔɪvstʌʊ̃w	Cov. “fivestones (when)”	ĩðəˈfɔm	Brown, SSB “in the form”

4.3. Onset changes

a. ð-reduction

ˈɑɬləˈtʰa:m	Stkpt. “all the time”	ɪn̪n̪iz	SSB. “in these”
ən̪n̪æʔs	ShB “and that’s”	kɔ:ɬɬəm	Psmsh. “call them”
ˈwən̪n̪i	Nor. “when the”	ˈən̪ˈn̪aʕ	Ed. “and that (was)”
in̪n̪ə	Cov. “in the”	ɑ̈ɬɬəz	Cov. “well, there’s”

b. h-dropping

ˈsinɚ Am "seen her" ˈætsɨ ż Am. "that's his"

c. 'palatalisation'

eʔkɑʃ e̊	Stkpt. "it costs you"	h ʧ	Am. "hit you"
ɹʊu ndʒə	ShB. "ruined your"	əʒ jʊuʒʊ	Psmsh. "as usual"
ʧə sɛɬf	Psmsh. "(mix) it yourself"	faɪndʒɔ	SSB. "find your"
didn̥ʧə	Cov. "didn't you"	wɔnʃjüud	Nor. "once you'd"

While each of these changes is relatively mild, large differences from citation form can be produced when they occur in combination, such as in [ˈmãʊ̃ʔn̥] for "mountain".

5. Once and future research

Naturalness in synthetic speech is an ongoing concern, especially with respect to prosody and emotion (SHIH *et al.*, [18], SCHRODER *et al.*, [17]) but also including style (TERKEN, [19]). It has been shown that casual speech forms can be generated using nonsegmental synthesis (COLEMAN, 1995), but it is not clear to what extent Coleman's insights are being used, and in general little work on including speech shortcuts can be found in the literature.

Gotthardson, (2005) has tested the idea that synthesised speech sounds more natural at faster rates in Swedish if it contains casual speech reductions. A further hypothesis was that more frequent words would sound more natural with greater reduction. Her set of Swedish casual speech rules was taken from work by JANDE [6, 7].

She synthesised a set of sentences spoken at a range of rates, a range of degrees of reduction, and a range of word frequencies and asked subjects to judge the degree of naturalness of the result by indicating on a scale from 'low' to 'high'. Results were not as expected: instead of finding the faster/more frequent forms to be more natural when phonologically reduced, most subjects preferred the canonical forms in all environents. A smaller group preferred the reduced forms in all environments.

It is promising that research along these lines exists, but there are several aspects of Gotthardsson's experimental design which I would question. In English, for example, I have not been able to find a strong correlation between increased speech rate and phonological reduction, i.e. the term 'fast speech phonology' is misleading for English. It would be useful to know whether this correlation is actually present in Swedish before assuming it to be so. My assumption is that *any* connected speech is closer to the real thing when it includes casual shortcuts than when it does not. Second, the notion that more frequent words show greater phonological reduction must be tested further: the most frequent words in English, for example, share features other than frequency. They tend to be short, closed-class words which are predictable or redundant and these factors probably conspire in their reduction. A similar but not identical claim, that subsequent

mentions of a noun are more phonetically-reduced (less peripherally articulated) after the first time in a discourse does not apply to phonology: Sotillo, (1997) has shown that whereas phonetic effects are sensitive to previous mention, phonological reductions are not.

My plan for English is to use tests used previously constructed for comparing synthetic speech with natural speech, i.e. to ask 1) is synthetic speech which uses native shortcuts easier or harder to understand? This can be tested by asking subjects to write or repeat what they hear rather than judging naturalness. Subjects can also be asked to choose between responses, i.e. "Did you hear X or Y"? 2) is synthetic speech containing reductions easier to remember than that without? This can be tested by asking subjects for summaries of stories heard days, weeks, or months before in the two styles, citation form and reduced.

Another reason for Gotthardsson's negative results may have been that she was using diphone synthesis to produce her reduced forms. Taking a hypothetical case for English, to generate [tɛs:] for 'tests', you would probably use diphones 'te' + 'es' + extra length for the second half of the second segment (possibly an 's+s' diphone). But typically in this case, a partial gesture towards the second 't' causes a loss of amplitude in the middle part of the frication. This is easy to model using terminal analogue synthesis or articulatory synthesis. To get it correct using diphones, you would need diphones 's + partially closed t' followed by 'partially closed t + s'. There would also be difficulty in generating English [kɑ̃ʔ] for 'can't', as there would not normally be diphones for nasalised vowels without accompanying nasals. Many diphone systems would probably not include glottal stop either, as it is not phonemic in English. In brief, many more diphones than are normally stored would be needed to generate English casual speech forms, maybe even prohibitively more.

References

[1] BARD E. G., SHILLCOCK R. C., ALTMANN G. T. M., *The recognition of words after their acoustic offsets in spontaneous speech: effects of subsequent context*, Perception and Psychophysics, **44**, 395–408 (1988).

[2] DENES P., *Effect of duration on the perception of voicing*, Journal of the Acoustical Society of America, **27**, 761–764 (1955).

[3] GROSJEAN F., *The recognition of words after their acoustic offset: evidence and implications*, Perception and Psychophysics, **38**, 299–310 (1985).

[4] HAWKINS S., HEID S., HOUSE J., HUCKVALE M., *Assessment of naturalness in the ProSynth speech synthesis project*, IEE colloquium on Speech Synthesis, London 2000, available at www.phon.ucl.ac.uk/home/mark/papers/iee00hawkins.pdf

[5] HOUSE A., WILLIAMS C., HECKER M., KRYTER K., *Articulation testing methods: Consonantal differentiation with a closed response set*, Journal of the Acoustical Society of America, **37**, 158–166 (1965).

[6] JANDE P.-A., *Evaluating rules for phonological reduction in Swedish*, Proceedings of Fonetik, pp. 149–152, 2003.

[7] JANDE P.-A., *Phonological reduction in Swedish*, Proceedings of the International Congress of Phonetic Science, 2003, pp. 2557–2560.

[8] JEKOSCH U., *Speech quality assessment and evaluation*, [in:] *Eurospeech '93*, Proceedings of the Third European Conference on Speech Communication and Technology, Berlin, September 1993, European Speech Communication Association, pp. 1387–1394, 1993).

[9] KRAFT V., PORTELE T., *Quality of five German speech synthesis systems*, Acta Acustica, 3, 351–365 (1995).

[10] LAURES J. S., WEISMER G., *The effects of a flattened fundamental frequency on intelligibility at sentence level*, Journal of Speech, Language, and Hearing Research, **42**, 1148–1156 (1999).

[11] LODGE K., *Studies in the Phonology of Colloquial English*, Croon Helm, 1984.

[12] LUCE P., *Neighborhoods of words in the mental lexicon*, Research on speech perception technical report no. 6, Indiana University, 1986.

[13] LIBERMAN A. M., DELATTRE P., COOPER F. S., GERSTMAN L. J., *The Role of Consonant-Vowel Transitions in the Perception of the Stop and Nasal Consonants*, Psychological Monographs, **68**, 1–13 (1954).

[14] MCCARTHY J., PRINCE A., *The emergence of the unmarked: optimality in prosodic morphology*, [in:] M. Gonzalez [Ed.], Proceeding of the North East Linguistic Society, **24**, 333–379 (1994).

[15] PISONI D., *Perception of synthetic speech*, [in:] van Santen, Sproat, Olive, and Hirschberg [eds.], *Progress in Speech Synthesis*, Springer, pp. 541–560, 1997.

[16] POLS L. C. W. *et al*., *Multi-lingual synthesis evaluation methods*, Proceedings of the 1992 International Conference on Spoken Language Processing, volume 1, pp. 181–184, Banff, Alberta, Canada, October 1992, University of Alberta, 1992.

[17] SCHRODER M., COWIE R., DOUGLAS-COWIE E., WESTERDIJK M., GIELEN S., *Acoustic Correlates of Emotion Dimensions in View of Speech Synthesis*, Proceedings of Eurospeech 2001, pp. 87-90, Aalborg, Denmark 2001.

[18] SHIH C., KOCHANSKI G. P., *Synthesis of prosodic styles*, 4th ISCA Tutorial and Research Workshop on Speech Synthesis, Scotland 2001.

[19] TERKEN J., *Variability and speaking styles in speech synthesis*, [in:] Keller E., Bailly G., Monaghan A., Terken J., Huckvale M. [Eds.], *Improvements in Speech Synthesis. Cost 258: The Naturalness of Synthetic Speech*, pp. 199-203, John Wiley & Sons, Chichester 2002.

[20] SHOCKEY L., *Sound Patterns of Spoken English*, Blackwell 2003

[21] VOIERS W., SHARPLEY A., HEHMSOTH C., *Research on diagnostic evaluation of speech intelligibility*, Research Report AFCRL-72-0694, Air Force Cambridge Research Laboratories, Bedford, Massachusetts 1975.

[22] WELLS, *Accents of English*, Cambridge University Press, 1982