The Role of Phonetics in Synthesis

Jan P. H. van Santen

Center for Spoken Language Understanding
OGI School of Science & Engineering at OHSU

ABSTRACT

This paper addresses three interrelated, broad questions: (i) What type of phonetic knowledge is used in text-to-speech synthesis? (ii) What good does it do? (iii) What future phonetics research does synthesis need? We argue that, depending on the specific architecture and aim of the system (i.e., open domain or closed domain), text-to-speech synthesis systems can incorporate a great variety of facts about human language. These facts do not necessarily take the form of manually crafted rule systems. These rules systems have often have been faulted for fragility, which in turn has been used as an argument for doing away with the incorporation of phonetic knowledge and using machine learning instead. The key importance of incorporating phonetic knowledge is its domain-independence, which lessens the dependence of a system's performance on the non-generalizable peculiarities of a training corpus. Moreover, current phonetic knowledge is not enough – needed improvements of speech synthesis quality need answers to many general phonetic questions; some examples of these will be provided. For this, however, closer cooperation is needed between the speech technology and phonetics communities.

1 INTRODUCTION

When discussing the use of phonetics in speech technology, one often has the impression that, on the one hand, members of the speech technology community view progress made over the last two decades as a process in which reliance on phonetic knowledge has been replaced by usage of machine learning algorithms in conjunction with speech corpora, while, on the other hand, members of the phonetics community criticize speech technologists for being linguistic ignorants and observe with glee the many shortcomings of current speech technology that allegedly are the result of this ignorance. This paper contends that both points are right. That is, no progress would have been possible without the revolutionary advances made in machine learning and availability of speech corpora. But also, ignoring phonetics – in particular, phonetics in the broad sense – may hamper further progress.

The first point hardly needs further arguments; we will not provide these. The second point will be elaborated upon in two ways. First, we demonstrate the *variety* of types of knowledge incorporated in speech technology. Specifically, we argue that phonetic knowledge incorporated in synthesis systems goes well beyond manually crafted rules. Second, we argue for the critical role of phonetic knowledge for domain-independent synthesis.

The structure of this paper is as follows. First, we provide a brief overview of speech synthesis. Next, the role of phonetics in synthesis is illustrated with the Bell Labs multi-lingual text-to-speech synthesis system [1]. This system is selected primarily because in its 2-decade history it has absorbed research of a large, multi-disciplinary group of scientists working at all levels of the system, and thus provides a unique opportunity for observing the role of knowledge in synthesis. Finally, we discuss research issues that are phonetic in nature and whose resolution could benefit further progress in speech synthesis.

2 SPEECH SYNTHESIS

2.1 Speech Synthesis: Overview

Most current systems employ the following multi-step process. *Text analysis* modules compute from text symbolic representations, such as phoneme labels, prosodic tags, or part-of-speech tags. Next, *Target prosody* modules compute timing information, pitch contours, and optionally also amplitude, spectral balance, and voicing characteristics. Finally, a *speech corpus* is searched for appropriate fragments of recorded speech (*acoustic units*), these units are *concatenated*, and optionally *modified* so that the output speech exhibits the pre-computed target prosody.

Systems differ primarily in the presence and role of target prosody computation and in the type of signal modification performed on the stored speech. In *unit table based* systems (e.g., the systems from Bell Labs, Elan Informatique, and SVOX), the speech corpus consists of a set of pre-excised acoustic units, mostly diphones; thus, no real search is needed, merely retrieval. At run time, the appropriate units are retrieved from the table, concatenated, and modified to produce the desired target prosody. This modification is extensive, because it involves both time warping and changing the pitch contour. Over the decades, a major focus

ISBN 1-876346-48-5 © 2003 UAB

of speech technology research has been on developing algorithms to perform modification operations with a minimum of signal distortion [2, 3, 4, 5, 6].

Newer systems (*unit selection* systems) capitalize on progress in storage capabilities, processing power, and search algorithms, by using substantially larger speech corpora. CHATR [7] went to the extreme of using *no* pitch modification, under the assumption (the *complete coverage* assumption) that for any sequence of target phonemes and target prosodic tags a corresponding sequence of speech intervals could be found. The advantage is that no signal modification is needed, thereby reducing signal distortion. The disadvantage is that, except for small domains, the complete coverage assumption is incorrect; this has dreadful effects on speech quality, because many units will be inappropriate or have spectral or pitch discontinuities. Post-CHATR unit selection systems have improved performance by using much larger and more carefully constructed speech corpora (e.g., [8]) or by using modest amounts of signal modification [9, 10].

2.2 Role of domain restriction in synthesis

The unrestricted domain consisting of all text materials in a given natural language contains an extremely large number of combinations (or units) of phone sequences and prosodic contexts; moreover, one cannot satisfactorily cover these units in a speech corpus by only focusing on frequent units [11, 12], for two reasons. First, the probability mass of "rare events" is near-certainty. Concretely, this means that almost any input sentence will require units that are not in any reasonably-sized speech corpus. Second, frequency distributions differ substantially between text types.

The combinatorial problems faced at the level of speech corpora also apply to training corpora used for other system components, e.g., the duration component.

In summary, systems differ radically in terms of their abilities to generalize beyond a speech corpus. Unit selection based systems may be virtually indistinguishable from natural speech for input sentences from, or highly similar to, those in the corpus, but may break down badly for other input sentences because they are based on the full coverage assumption. Traditional systems fall in between: They never sound natural but also do not break down.

The critical role of phonetic knowledge is that it is *general*, and not tied to a specific speech corpus. We now discuss how such knowledge is, or can be, incorporated in synthesis systems.

3 THE BELL LABS SYSTEM

Prior to its complete rewrite as a prelude to commercialization in the mid-90's, the primary goal for th Bell Labs system was more that of a test bed for research than a commercial TTS engine. This makes the system unusual (probably the only other system similar in this respect is Dectalk and its pre-cursors [13]), but also useful for the present paper.

3.1 Text Analysis

Its Text Analysis component uses a mixture of knowledge based and statistical approaches: Dictionary lookup, parts-of-speech tagging, syllabification rules that are based on the sonority hierarchy, morphological analysis based on general linguistic theories about morphology, heuristic pronunciation rules, statistically trained algorithms for word segmentation (Chinese), homograph disambiguation, accent assignment, and phrase break assignment. In addition, knowledge also enters the choice of data features used for the statistically trained methods. For example, [14] used in their statistical method a mixture of features, including parts of speech, distance from and to punctuation, and lexical items. Selection of these features was based on psycholinguistic research.

Knowledge also enters on a general, "architectural", level. The current version of the system is based on weighted finite state transducers (WFST's). This system was created to address problems posed by a variety of languages, including Chinese, Japanese, and Russian, that were not readily solvable in standard architectures. For example, in Russian "text normalization" (e.g., for pronouncing the "%" sign) requires non-trivial linguistic analysis and hence cannot be handled via a pre-processor, as is commonly done in standard architectures. WFST's allow for a non-sequential decision making process in which multiple constraints are combined and jointly optimized. The realization that this was necessary is an example of how even system architecture can reflect linguistic knowledge.

3.2 Duration

The Bell Labs system uses "sum-of-products" models for predicting duration [15, 16]. This is a family of simple equations that can be considered a generalization of the additive and multiplicative models. While mathematically trivial, these models incorporate important facts about duration. First, they state that most factors affecting duration are *directionally invariant*, i.e., holding all else constant, stressed is longer than unstressed, and phrase-final is longer than phrase-medial. Second, substantial knowledge entered into the decision which contextual factors to include and which sum-of-products to use. For example, it is a general fact about American English that vowels are longer when followed by a voiced consonant. and that this effect is much larger pre-pausally [17]. This can be incorporated by using a sum-of-products of the form:

$$\mathrm{DUR}(vowel\ c, stress\ s, position\ p, voicing\ V) = s_1(v) + s_2(s) + s_3(p) \times s_4(V) \quad (1)$$

Third, since different phoneme classes occur in different contexts, and are also differentially affected by the same factors (e.g., consonants in syllable codas are not lengthened by stress but consonants in onsets are), phoneme classes must be categorized, again requiring phonetic knowledge.

It is important to realize that these decisions are based on *systematic phonetic experiments*, and not on cursory inspection of the literature on duration. The focus of these experiments is not that of estimating parameters (e.g., $s_1(v)$, $s_2(s)$, $s_3(p)$), but that of systematically varying factors and observing their interactions. For example, Klatt [17] conducted a study in which most factors were kept constant (e.g., sentence length), and provided a compelling argument that the interaction patterns found were not a peculiarity of the "training corpus" he analyzed but represented a general fact about American English speech.

Of course, with enough data, and using powerful statistical methods [15], these decisions could in theory have been derived automatically from the training data. However, the actual amount of training data is likely to be enormous, and one loses the benefit of a connection with an established body of phonetic research knowledge – since it is there, why not use it?

3.3 Intonation

The intonation component is based on the superpositional model according to which a pitch curve can be written as the sum of simpler component curves: a phrase curve associated with phrases, accent curves associated with pitch accents, and segmental perturbation curves associated with individual phonetic segments. Admittedly, the superpositional concept is controversial [18], but that is not the point. The point is that this concept is based on knowledge, including hypotheses about quasi-independent processes in the vocal chords [19] and analyses of production data [20]. Further choices in the implementation, such as on which factors parameters depend, are also based on knowledge from systematic phonetic experiments.

3.4 Signal Processing

Finally, although the signal processing component is concatenative and hence is not based on the detailed level of articulatory or acoustic modeling of, for example, MITalk [13], it nevertheless can be said to use knowledge [2, 21]. First, there are language dependent facts (e.g., in American English – but not in UK English – one needs consonant-vowel-r triphones because of the strong coarticulatory effects of [r]; Italian trill [22]; aspirated voiced stops in Hindi; vowel-devoicing in Japanese). Second, there are several details in the signal processing operations that reflect knowledge. For example, the temporal compression/stretching operation stretches out primarily the central portion of vowels instead of the initial and final portions. This is based on studies on vowel lengthening by, e.g., [23]. Third, further details are based on perceptual studies showing that certain distinctions need or do not need to be made (e.g., one cannot hear the difference between [s-t]+[t-o] and [s-p]+[t-o]; in other words, the [s] does not audibly depend on the place of articulation of the subsequent stop.)

3.5 Knowledge used in the Bell Labs System

The Bell Labs System incorporates many different types of knowledge. It should be emphasized, however, that almost all components of this system are at least partially data driven. In this sense, it is fundamentally different from the earlier MITalk system [13] in which every component consisted of manually constructed rules and manually adjusted parameter values.

4 NEEDED PHONETICS RESEARCH

Research is needed on several issues. A random selection from these:
1. Perception of spectral discontinuities of the type that occur in concatenative synthesis.
2. Perception of discontinuities in intonation contours.
3. Understanding sub-segmental timing in speech production.
4. Mimicking vowel reduction.
5. Understanding variability in speech production – both inter- and intra-speaker variability.
6. Acoustic invariances of perceptually equivalent pitch contours.
7. Multidimensional modeling of all acoustic prosodic features – F_0, local acceleration, spectral balance, loudness, etc.
8. How to measure the impact of emotional speech on listeners.

5 CONCLUSIONS

Text-to-Speech technology uses more linguistic knowledge, structures, and concepts than is generally realized. We believe that speech technology could benefit by incorporating even more knowledge. Some of this knowledge is already available, but there is a large list of phonetics research questions that are currently not addressed by the phonetics community. At the same time, the incorporation of linguistic knowledge and concepts requires speech technology systems to have receptive architectures.

Neither this research nor these architectures will materialize unless organizational and educational bridges are built between these two fields of research. The recommendation is obvious and has been made by others (e.g., [24]): we need to enhance growth of a "bridge

ISBN 1-876346-48-5 © 2003 UAB

field" in the form of *mathematical* or *computational phonetics*. It would not be the first time in the history of science that progress is hampered by the presence of high walls between narrowly defined disciplines.

ACKNOWLEDGMENTS

This paper was written with the partial support of the National Science Foundation, award numbers 0205731, 0117911, and 0082718. The paper has benefitted from discussions with Gilead Cohen, Peter Heeman, Hynek Hermansky, John-Paul Hosom, Alexander Kain, Goh Kawai, Esther Klabbers, Taniya Mishra, Xiaochuan Niu, Richard Sproat, and Johan Wouters.

REFERENCES

[1] R. Sproat, *Multilingual Text-to-Speech Synthesis: The Bell Labs Approach*, Kluwer, Boston, MA, 1997.

[2] J. Olive and M.Y. Liberman, "Text to speech – an overview," *Journal of the Acoustic Society of America, Suppl. 1*, vol. 78, no. Fall, pp. s6, 1985.

[3] F. Charpentier and E. Moulines, "Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones," in *Proc. of Eurospeech-1989*, Paris, 1989, pp. 13–19.

[4] Y. Stylianou, *Harmonic plus noise models for speech, combined with stat. methods for speech and speaker modification*, Ph.D. thesis, ENST, January 1996.

[5] G. Richard and R. d'Allessandro, "Modification of the aperiodic component of speech signals for synthesis," in *Progress in speech synthesis*, J. van Santen, R.W. Sproat, J. Olive, and J. Hirschberg, Eds., pp. 41–56. Springer, 1996.

[6] H. Kawahara, I. Masuda-Katsuse, and A. de Cheveign, "Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based f0 extraction: Possible role of a repetitive structure in sounds," *Speech Communication*, vol. 27, pp. 187–207, 1999.

[7] A.W. Black and P. Taylor, "CHATR: a generic speech synthesis system," in *Proc. COLING94*, 1994, pp. 983–986.

[8] M. Beutnagel, A. Conkie, J. Schroeter, Y. Stylianou, and A. Syrdal, "The AT&T NextGen system," in *Proceedings of the Joint Meeting of ASA, EAA, and DAGA, Berlin, Germany, March 15–19, 1999*, Berlin, Germany, 1999.

[9] R.E. Donovan, *Trainable speech synthesis*, Ph.D. thesis, Cambridge University, 1996.

[10] A.P. Breen and P. Jackson, "Non-uniform unit selection and the similarity metric within BT's Laureate TTS system," in *Third ESCAWorkshop on speech synthesis*, Jenolan Caves, Australia, 1998.

[11] J. van Santen, "Combinatorial issues in text-to-speech synthesis," in *Proceedings Eurospeech-1997*, Rhodos, Greece, 1997.

[12] B. Moebius, "Rare events and closed domains: Two delicate concepts in speech synthesis," in *Workshop on Speech Synthesis*, Pilochry, 2001, ESCA.

[13] J. Allen, S. Hunnicut, and D.H. Klatt, *From text to speech: The MITalk System*, Cambridge University Press, Cambridge, U.K., 1987.

[14] M.Q. Wang and J. Hirschberg, "Automatic classification of intonational phrase boundaries," *Computer Speech and Language*, vol. 6, pp. 175–196, 1992.

[15] J. van Santen, "Analyzing N-way tables with sums-of-products models," *Journal of Mathematical Psychology*, vol. 37, no. 3, pp. 327–371, 1993.

[16] J. van Santen, "Assignment of segmental duration in text-to-speech synthesis," *Computer Speech and Language*, vol. 8, pp. 95–128, April 1994.

[17] D.H. Klatt, "Interaction between two factors that influence vowel duration," *J. Acoust. Soc. Am.*, vol. 54, pp. 1102–1104, 1973.

[18] D.R. Ladd, *Intonational phonology*, Cambridge University Press, Cambridge, UK, 1996.

[19] H. Fujisaki, "A note on the physiological and physical basis for the phrase and accent components in the voice fundamental frequency contour," in *Vocal physiology: voice production, mechanisms and functions*, Fujimura, Ed. Raven, New York, 1988.

[20] J. van Santen and B. Möbius, "A model of fundamental frequency contour alignment," in *Intonation: Analysis, Modelling and Technology*, A. Botinis, Ed. Cambridge University Press, 1999, In press.

[21] J. Olive, "A new algorithm for a concatenative speech synthesis system using an augmented acoustic inventory of speech sounds," in *Workshop on speech synthesis*, Autrans France, 1990, ESCA, pp. 25–30.

[22] Chilin Shih, "Synthesis of trill," in *Proceedings of the International Conference on Spoken Language Processing*, Philadelphia, 1996, ICSLP, pp. 2223–2226.

[23] Th. Gay, "Effect of speaking rate on diphthong formant movements," *Journal of the Acoustical Society of America*, vol. 44, pp. 1570–1573, 1968.

[24] R. Moore, "Computational phonetics," in *Proceedings of the 13th International Congress of Phonetic Sciences*, Stockholm, 1995, vol. 2, pp. 68–71.