No future for comprehensive models of intonation?

Gérard Bailly

February 6, 1996

The general problem of speech synthesis is to produce the most intelligible and natural audio (and visual) stimuli from a set of linguistic, paralinguistic or emotional instructions. While traditional text-to-speech systems are intrinsically limited by the automatic processing of text to the rendering of lexical, syntactic and some semantic information, the insertion of speech synthesis in more sophisticated person-machine communication systems - including person-person interpreted dialogue - requires the generation of far variable speaking styles than so far studied. The prosodic generators described in this chapter all use automatic learning: the prosodic parameters most likely to be associated with the instructions mentioned above are generated by an associator. This associator, whether regression tree, multinomial regression, neural network...- is learned through examples of natural associations. Does this mean that prosodic research will be restricted to only identifying the most appropriate instruction descriptors and should therefore focus only on labelling ?

1 Learning from examples

Empirical studies share common methodological problems. We will attempt to elucidate some of these problems with reference to some points made in the following papers.

1.1 The reference corpus

The first methodological problem is how to obtain relevant training material. A speech segment is subject to the interaction of five main components: (a) motor control and production constraints, often referred to as coarticulation effects, (b) perceptual constraints, (c) structure of the language, (d) intrinsic environmental conditions - speaker's commitment, mood and listener's expectations, and (e) extrinsic environmental conditions - such as signal-to-noise ratio. The essential feature of speech is its ability to adapt to these constraints in order to ensure the most efficient - economical in all terms - communication. In order for a model to adequately learn the strategy followed by a particular speaker in a certain situation, the reference corpus must have a sufficient statistical coverage of the constraints mentioned above. As mentioned by Jan van Santen, an unconstrained recording of spontaneous speech will often result in a corpus with severe sparcity of factors. The lexicon daily used by speakers is often limited to a few thousand words: in the phonetic dictionary for French "De A à Zut" [5] published in 1992, 102137 lexical items and 304752 sounds from various sets of conversations have been segmented by hand. The number of occurrences `N` for each word is a drastic function of its rank R: `N = 10.75 - 1.25 log(R)`, e.g. the most frequent word "De" (of) occurs 3305 times whereas the word "appeler" (to call) with rank 826 only occurs 10 times, and more than 4000 words only occur once. Three solutions may face this sparcity: (a) introducing experimental procedures to directly or indirectly control the speaker's production: from read to real spontaneous speech the span of possible experimental designs is large: Black for example analyses the speech produced by the agent in multi-modal dialogues between an agent and clients; (b) using the generalisation abilities of statistical models such as regression trees (see also Hirai et al), neural networks (see Fujio et al) or multinomial regressions (see van Santen); or (c) using models which can be trained separately in order to avoid the combinatory explosion (see the use of a superposition model in Hirai et al and the criticism of two-stage models for duration control in van Santen). But the problem of sparcity is perhaps ill-posed, as put forward by Campbell. If the reference corpus reflects the intended characteristics of the target speech, then the phonetic balance represented in the corpus is adapted to the degrees-of-freedom imposed by the application: the synthesis can thus been considered as an information retrieval process which selects the reference speech segments that most closely match the target characteristics. This argument is completed by naturalness constraints: concatenative synthesis outperforms complex rule-generated synthetic speech because prosodic deformation of natural samples does not introduce disruptive artifacts such as those introduced by speech production modelling or stylisation procedures. The best way to preserve the intrinsic naturalness of the sounds of the reference database is to leave most parts of the speech untouched.

1.2 Labelling the corpus

Once the database has been collected according to some given experimental settings, the audio-visual signals and the hypothesised constraints which the speaker obeyed during his oral production have to be explicitly associated in the reference database. These symbolic descriptions range from subjective general description of the emotional state of the speaker, and environmental conditions, to phonological descriptions closely linked with the actual realizations. The limit between the descriptions of the intended speech acts and their actual realization is fuzzy and often considered as information best expressed in terms of features - thus subject to interpretation - vs measurable variables.

1.2.1 a case study: the pause

Pause insertion is one such phenomenon, which is often treated at both symbolic and numerical levels: pause insertion is often considered as part of the phonological description - *i.e.,* the pause is treated as part of the phonetic string, delimits and determines largely phonological constituents, and is predicted from linguistic structure (see Fujio et al). Whereas the absence/presence of a pause is then used as a feature to determine duration and melodic contours, the generation of its own duration is often not considered with the same care.

However, the work done by Grosjean and colleagues [10, 14] has shown that pause durations encode fine details of the linguistic structure and suggests that the pause and the rhyme of the preceding syllable function as a coherent rhythmical unit (see also the similarities between pause and prosodic phrase boundary locations in Fujio et al). A computational model has been proposed in [3], showing that it is possible to compute both pause insertion and duration from boundary strength and speech rate without any featural interface. This example shows that not only the choice and ordering of prosodic predictors may affect the performance of automatic learning but also the judicious use of theoretical models of prosodic structure.

1.2.2 Discourse structure

Campbell's and Black's papers in this chapter are deeply concerned with the nature of high level discourse information that could be relevant for describing speaker-listener interactions. The advantage of such synthesis applications is their intrinsic rich description of the synthesis task: in order to generate an adequate part-of-speech the dialog system must have a complete and accurate view of the current state of the dialog. How far do the features predicted by the dialog system describe the actual intentions of a real speaker ? Each aspect of the speech enriches the common knowledge the speaker and the listener in order to further the purpose of the interaction. The prosodic structure may only carry a small part of this information and the speaker makes at each moment important choices according to what is most relevant for him and what he supposes the listener already knows and has understood. Thus analysis of such dialogs should face sparcity, underspecification and uncompleteness.

For example, given/new information not only influences accents but also may result in more global behaviour. We recorded recently [15] a large corpus of sentences pronounced with different attitudes. Incredulous questions where the speaker repeats some utterance just heard often miss syntactically related prosody. This information has been already captured and the speaker only wants confirmation on the entire assertion.

1.3 The sub-symbolic paradigm: training an associator

Once the reference corpus has been designed, recorded and labelled, diverse learning paradigms have been proposed, including CART techniques, multinomial interpolators or neural networks. These tools share a common feature: they produce a least-square fit of reference numerical values with predicted values. As previously argued by van Santen [23], such an optimisation procedure should be preceded by a selection of features affecting particular subsets of sounds. Following this procedure, prosodic trajectories are thus predicted on a sound-by-sound basis. The cohesion of the temporal structure of speech events and the kinematics of the trajectories is thus only maintained through the phonological relationship these sounds entertain with higher levels. When phonological and phonotactic information is given with sufficient precision, such systems might be able to give an accurate prediction. However, since prediction errors are rarely null, it should be perhaps interesting to evaluate the perceptual relevance of these errors. The appropriate answer to this question should be to incorporate more perceptual knowledge in the processing of reference prosodic data *e.g.,* various intonation models for automatically simplifying the melodic curve have been proposed to date [21, 7] including the one used previously by Campbell [12]. Similarly, we may question whether sound durations contribute equally to the perception of momentary tempo. For example, perceptual experiments described in [2] showed that listeners are more sensitive to errors in relative timing of vocalic onsets than in errors uniformly distributed among segments.

1.4 The morphological paradigm

These functional models of intonation may adequately describe the prosody used by reference speakers. They do not however supply any comprehensive model of the speaker-listener interaction. One could claim that this lack of theoretical background is the key of the success of sub-symbolic approaches in speech technology. Synthetic speech may have prosody as planes have wings: wings assume some common function but need not be as rapidly moving as those of birds (although the astute reader may argue that birds do not need airports and can land on electric cables). In the present case prosody is charged with carrying information on the human's concrete and abstract world: prosodic patterns such as pitch accents, phrase accents and boundary tones are phonological objects on which humans belonging to a certain linguistic community have agreed. It is thus quite legitimate to strongly base our prosodic models on the way people do use prosody to encode "what they think and how they feel about what and when they say" [6]. Our speech communication system is robust because it is redundant: coarticulation, and anticipatory patterns are effectively used by the perceptual system. Each sound thus contributes at more than one prosodic level; all melodic descriptions include at least a *phrase* level, to control utterance chunking, as well as signalling *word* accents. How many such prosodic elements overlap in time? What are their phonetic characteristics and how do they combine ? These questions are still open: however some morphological decomposition of the F0 curve have already been proposed: Thorsen [22] proposes a model of Danish intonation where word accent are added to an Overlap-and-Add of declination lines. A similar approach was also proposed by O'Shaughnessy and Allen for English [18], Gårding for Swedish [11] and of course by Fujisaki [9] for Japanese. A more generic approach involving an Overlapp-and-Add of intonation contours has been proposed by Aubergé [1] and is actually further developed [15].

2 Conclusions

The experimental approaches put forward in this chapter make claims for three main research efforts.

First, both the use of statistical methods to automatically learn phonetic models and the search for morphological descriptors claim for the creation of large prosodically labelled corpora of speech data. The use of phonological descriptors should not obscure the concrete prosodic phonemena and should be used as filters to enable speech researchers to extract lawful variability from prosodic signals: discourse intonation not only encodes salience and segmentation but also more subtle attitudes and affects with different *illocutory forces*, and the predominance of 'prominence' in current prosodic phonology may hide more global patterns.

Secondly, much care should also be taken when defining and designing these research corpora: the development of experimental physics at the beginning of the 18th century demonstrated that the ultimate structure of nature may be captured and clever ideas validated only by using well-controlled experiments. Recording predefined text with different attitudes or emotions [16] while reading or by memorising it, or describing figures appearing on a computer screen [20] are just two examples of such research paradigms, many more experimental settings have to be invented.

Thirdly, as the number of instructions which the synthetic prosodic generator should obey increases, the combinatory explosion of interacting factors will have to be mastered. The coherence and the statistical significance of the data needed to train associators or to store prototypical templates will be difficult to maintain as long as comprehensive models of rhythm and intonation are not able to propose phonetic models where segments and prosody can be processed in parallel [19] and where discourse structure and phrasal levels are well identified. Superpositional models which combine prosodic contours with specific operators have already been used in synthesis [9] as well as in recognition experiments [17, 13]. Linear models combined with non-overlapping prototypical prosodic templates have the advantage of simplifing the composition/decomposition problem and may drastically limit the amount of learning data required: user-specific templates may be learned separately for each representation level (discourse, turn, paragraph, sentence, proposition, group ...) as microprosody contours are often added to a melody free of segmental effects.

I claim that a better understanding of the phonetic representation used to encode the various functions of prosody are the right tools to understand strategies used by speakers even within spontaneous corpora. Psycholinguistic experiments often use constraints or perturbations together with control experiments to examine the robustness of human biological systems. Arguments against the use of carefully controlled data always focus on differences between read speech and freer speaking styles [4]. We should also focus on similarities: linguistic and prosodic strategies often preserve local syntactic or prosodic structures as already observed by Fónagy and others:

"The opinion that any utterance is an instantaneous production ... applies only to the domain of modern poetry. Except for this limited domain, our utterances are made of large prestored pieces..." (Translated from [8, p.182])

3 Bibliography

References

[1] Aubergé, V. Developing a structured lexicon for synthesis of prosody. In Bailly, G. and Benoît, C., editors, *Talking Machines: Theories, Models and Designs*, pages 307–321. Elsevier B.V., 1992.

[2] Barbosa, P. and Bailly, G. Characterisation of rythmic patterns for text-to-speech synthesis. *Speech Communication*, 15:127–137, 1994.

[3] Barbosa, P. and Bailly, G. Generation of pauses within the z-score model. In van Santen, J.P.H., Sproat, R.W., Olive, J.P., and Hirschberg, J., editors, *Progress in Speech Synthesis.* Springer Verlag, New York, in press.

[4] Blaauw, E. *On the perceptual classification of spontaneous and read speech* OTS Dissertation Series. ISBN 90-5434-045-2, 1995.

[5] Boë, L.J. and Tubach, J.P. *De A à Zut: dictionnaire phonétique du français parlé.* ELLUG - Université Stendhal, Grenoble - France, 1992.

[6] Bolinger, D. *Intonation and its uses.* Edward Arnold, London, 1989.

[7] d'Allessandro, C. and Mertens, P. Automatic pitch contour stylization using a model of tonal perception. *Computer Speech and Language*, 9:257–288, 1995.

[8] Fónagy, I., Bérard, E. and Fónagy, J.", Clichés mélodiques. *Folia Linguistica*, 17:153–185, 1984.

[9] Fujisaki, H. and Sudo, H. A generative model for the prosody of connected speech in japanase. *Annual Report of Engineering Research Institute*, 30:75–80, 1971.

[10] Gee, J.P. and Grosjean, F. Performance structures: a psycholinguistic and linguistic appraisal. *Cognitive Psychology*, 15:418–458, 1983.

[11] Gårding, E. Intonation parameters in production and perception. In *International Congress of Phonetic Sciences*, volume 1, pages 300–304, Aix-en-Provence, France, 1991.

[12] Hirst, D., Nicolas, P., and Espesser, R. Coding the f0 of a continuous text in french: an experimental approach. In *International Congress of Phonetic Sciences*, volume 5, pages 234–237, Aix-en-Provence, France, 1991.

[13] Jensen, U., Moore, R., Dalsgaard, P., and Lindberg, B. Modelling intonation contours at the phrase level using continuous density hiddeen Markov models. *Computer Speech and Language*, 8:247–260, 1994.

[14] Monnin, P. and Grosjean, F. Les structures de performance en français : caractérisation et prédiction. *L'Année Psychologique*, 93:9–30, 1993.

[15] Morlec, Y., Bailly, G., and Aubergé, V. Synthesis and evaluation of intonation with a superposition model. In *EUROSPEECH*, volume 3, pages 2043–2046, Madrid - Spain, 1995.

[16] Mozziconacci, S. Pitch variations and emotions in speech. In *International Congress of Phonetic Sciences*, volume 1, pages 178–181, Stockholm - Sweden, 1995.

[17] Nakai, M. and Shimodaira, H. Accent phrase segmentation by finding n-best sequences of pitch pattern templates. In *International Conference on Speech and Language Processing*, volume 1, pages 347–350, Yokohama, Japan, 1994.

[18] O'Shaughnessy, D. and Allen, J. Linguistic modality effects on fundamental frequency in speech. *Journal of the Acoustical Society of America*, 74(4):1155–1171, 1983.

[19] Pasdeloup, V., Morais, J., and Kolinsky, R. Are stress and phonemic string processed separately? evidence from speech illusions. In *Proceedings of the European Conference on Speech Communication and Technology*, volume 2, pages 775–778, Berlin, Germany, September 1993.

[20] Swerts, M.G.J. and Collier, R. On the controlled elicitation of spontaneous speech. *Speech Communication*, 11:463–468, 1992.

[21] ten Bosch, L. Algorithmic classification of pitch movement. *Working Papers of Lund University*, 41:242–245, 1993.

[22] Thorsen, N.G. Standard Danish sentence intonation – phonetic data and their representation. *Folia Linguistica*, 17:187–220, 1983.

[23] van Santen, J.P.H. Contextual effects on vowel duration. *Speech Communication*, 11:513–546, 1992.