# SYMPOSIUM 2: Units in Speech Synthesis

Chairman: *J. Allen, Cambridge, Mass., U.S.A.*

Panel members: *O. Fujimura, J. Holmes, R. Carlson, B. Granström, J. Pierrehumbert*

The goal of synthetic speech algorithms is to provide a means to produce a large (infinite) set of speech waveforms. The major questions posed to a designer of such a system include the question as to what input is to be used, by what set of transformations is the speech to be produced, and by what quality measures can success be judged. At a general level, the issues are the choice of *representations* and the provision of *transformations* between these representations. The complexity of the relation between input (e.g., text) and output (the speech waveform) forces the introduction of intermediate levels of representation. These levels of representation are arrayed in two hierarchies. First, there is the structural hierarchy, wherein bigger constructs are made from smaller ones. There is a feeling that over the last 20 years during which speech synthesis by rule has been studied, the level of rule complexity at any particular level is staying about the same. This means that in order to introduce additional constraints, greater modularity must be introduced into the overall speech synthesis system together with narrowly constrained interaction between these domains. By such a means the structural hierarchy becomes richer, but the level of complexity at any level of the hierarchy does not grow to an unreasonable level. The second kind of hierarchy is the qualitative structure that exhibits the differing nature of constraint domains that are needed to specify the output speech waveform. These constraint domains include phonetics, phonology, syntax, semantics, acoustics, anatomy, physiology, and computation.

In the following paragraphs, the desiderata for the choice of units for speech synthesis are discussed.

1. The large number of utterances that must be created by a speech synthesis system forces *composition* from a smaller number of basic units.
2. The larger the unit, the more items of each unit there are. Thus at the small end, there are relatively few phonemes, but as the units become larger through diphone, demisyllable, syllable, word, phrase; . . . the number of items of the unit grows without limit. Furthermore, the smaller the unit, the more abstract it is likely to be. There is a general tendency towards increased abstraction as additional knowledge about speech becomes codified.
3. The choice in representation of units must allow sufficient degrees of freedom to control all significant aspects of the waveform. Questions as to

what is significant must be answered through perceptual tests and there is a great need for increased investigations in this area.

4. It must be possible to *analyze* speech in terms of the selected units and their parameters. Clearly, as the units become more abstract, there is increasing difficulty in relating surface measurements to the abstract structural relationships. Nevertheless, there is an increasing trend to introduce these structures in order to be able to generate the rich variety of surface phonetic detail. This phenomenon can be readily observed in the area of suprasegmental control, where the complexity of specifying a wide variety of fundamental frequency contours requires both a rich structural framework and a substantial set of complementary procedures. Another reason for the introduction of abstract structural models is the need for automatic training. Large databases of speech waveforms must be analyzed in terms of these models in order to accurately characterize the needed control parameters. While substantial abstraction can be introduced in order to characterize generalizations in the underlying framework, these abstractions cannot become too distant from the surface phonetic reality if automatic training is to be possible. Furthermore, substantial improvements in speech synthesis quality are probably impossible without the introduction of semi-automatic techniques for improving the quality of the control parameters.
5. It must be possible to *compose* an utterance by interpretive processes on the chosen units. Here a tension between internal vs. external structure and cohesion can be observed. Thus, the phoneme has relatively little internal structure, but considerable external structure is needed to specify the transitions between phonemes at their boundaries. Diphones and syllables, on the other hand, have a richer internal structure, and require less specification at their boundaries. Advocates of larger units, such as the diphone and the demisyllable, maintain that 'hard' (or physiologically determined) coarticulation is captured internally within the unit, thus leading to simpler composition functions. On the other hand, there is a growing feeling that the units should be sufficiently rich internally to allow for the automatic centralization of vowels and durational adjustments without the necessity for these changes to be imposed entirely externally. The search for the 'natural joints' of language will certainly continue in an effort to determine the optimal units, but high quality synthetic speech probably requires the same amount of detailed knowledge of speech and language whether the knowledge is represented in terms of either internal unit complexity, or complexity in the external composition function.
6. Consideration of the *composition function* indicates that the chosen units can be either used in *compiled*, static form or *interpretively*. Thus the choice of units is related to the question as to whether perceptually significant knowledge of speech should be represented structurally in static form or procedurally in terms of interpretive rules. This contrast does not have to be viewed as a necessary tradeoff, since it is certainly possible to have increased complexity in terms of the structure of the units together with increasingly rich procedural interpretive processes. In fact, the choice of increasingly complex abstractions demands a correspondingly complex interpretive process to realize the complete variety of surface phonetic waveforms.

7. The *binding time* of decisions must be specified. Some units imply early binding time. Thus compiled structural lexical forms, such as diphones, imply less complicated procedures at composition time. Such a choice may imply substantial effort to obtain the original compiled structural forms, but less computation when the speech waveform is created. Alternatively, late binding implies greater flexibility to modify aspects of the waveform. The units of speech synthesis tend to be more abstract in this case, and there is more emphasis on procedures to compose the units. In terms of general tendencies among computational procedures in the large, there is a tendency towards late binding. Modern computational resources support this flexibility, but it demands comprehensive modeling and training procedures.
8. The units should have a *natural relation* to the waveform production model. There is still no agreement as to the choice of model for constraining the output speech waveform. Most schemes employ a source/filter model, but there are differences as to whether cascade, parallel, or other combinations of resonators are used. Furthermore, the source model varies substantially, and it is clear that a great deal of work is needed to improve this model. Articulatory models are used in some instances, but the relative lack of articulatory data has slowed progress in this area. Whatever model is selected, it is important that it be possible to readily compute the values of the control parameters from this model through relatively straightforward computations based on the speech waveform.
9. The units should provide *insightful* relevance to the research literature. For this reason many synthesis schemes rely on the specification of formant frequencies and bandwidths. When linear predictive coding is used, it is often necessary to transform back to the formant space in order to edit parametric representations for improved speech quality. It is possible that new more complex units and representational frameworks may provide good synthesis, new insight, and also stimulate new research. The tendency over time is likely to involve the development of increasingly complex unit frameworks which in turn imply new classes of perceptual experiments and training procedures which should both improve speech quality and provide greater understanding of underlying speech processes.
10. There are several important *levels* of units that are of varying *scope*. In general, all aspects of linguistic structure are reflected in the acoustic waveform. Thus there is no one unit which is more important than all others. The levels of structure currently recognized are: discourse, sen-

tence, clause, phrase, word, morpheme, metrical foot, syllable, phoneme, and feature. These have all been found useful in linguistic analysis for reasons of distribution and contrast, but they each exhibit some intrinsic cohesion and place focus on the units themselves. Thus each of these units implies an internal theory, together with a relation to other units and their corresponding theory. There is a tendency to think of these different levels of units with their attendant scopes as being arrayed over a temporal axis. While such a view provides a convenient geometric intuition, contemporary notions of abstraction may not rely on notions of temporal scope at all levels. Thus at the individual segment feature level the initial specification may include an unordered set of features which is then transformed at a higher level to derive temporal extent. Research in the asynchronous control of these features reveals the need for this varying scope at the feature level, as in the control of nasality.

11. The hierarchy of units should provide for a *gradual* progression through the various representations. That is, a sufficient number of units should be available so that the transformation from one level to another through the hierarchy is relatively small and direct. Thus for example, phonemes are transformed to allophones and then to targets, parameters and finally to the speech waveform. In this way, understanding of the speech process is modularized at each level, thus simplifying the rule structure at each level, even though the total number of rules over the entire system may be large.
12. The notion of *target* serves as a bridge between abstraction and physical properties. There is a tendency for the targets themselves to become complex objects exhibiting composition, scope, and internal cohesion. For example, fundamental frequency contours must be characterized at both the phrase level and in terms of 'segmental' effects. It is seen that the term 'target', originally motivated by the characterization of formant trajectories, is increasingly synonymous with 'abstract representational structure'. The role of *variability* in the choice of units must be considered in the design of any speech synthesis system. It is often felt that many attributes of these units are necessary, and hence admit of no variability. Recent understanding, however, indicates that many cues are redundant, and that some contextually determined integrative process that may be deterministic in nature determines the cues that must be instantiated in order to derive the intended percept. Thus there may be a level of variation on the surface that appears to be free or random, but which in fact is a highly complex manifestation of underlying regularity. There are no contemporary speech synthesis systems that begin to approach the level of surface phonetic variability observed in natural speech. Models are needed that characterize this variability in terms of redundant cues, idiosyncratic gestures associated with a given speaker, and true random variability. The level of understanding needed to provide this control is immense, and will doubtlessly elude practical systems for many years to come.

13. The units of speech synthesis should be able to support a wide class of *voices*. Most contemporary systems focus on the adult male voice, but child, and female voices must be obtainable by any general system, and there is a need for the ability to synthesize several languages and diverse dialects within the context of one overall synthesis framework. A few contemporary systems are capable of synthesizing utterances in several languages within the same framework, but these systems generally sacrifice quality of synthesis in any particular language for breadth of coverage over the diverse linguistic forms.
14. The choice of units is intimately related to questions of *intelligibility*, *naturalness*, and *reduction* phenomena. To date, most research in speech synthesis has aimed at producing intelligible speech, without as much attention being focused on issues of naturalness and vowel reduction. There is still much to be learned about vowel reduction, and some feeling that such reduction may come from redundancy when viewed in the context of the sum total of all constraints in force at any point in the utterance. There is a tension between intelligibility and naturalness in that if a system aims for substantial reduction in the name of naturalness, intelligibility may suffer. It is also very difficult to evaluate naturalness, although some success has been achieved using cognitive loading techniques. Nevertheless, this class of psycholinguistic experimentation is still in its infancy and a great deal of work needs to be performed. At the discourse level, there is increased need for naturalness since attributes of focus and old/new information cause substantial stressing and destressing over that which would be predicted for a sentence spoken in isolation. It should also be remarked that individual segment intelligibility still leaves much to be desired, and the codification within an insightful rule structure of the vast amount of segmental phonetic detail must be substantially improved if there is to be any appreciable gain in synthetic speech quality.
15. The units may exploit available technology, but they should not be overly determined by it. Computational technology typically involves space/time tradeoffs. Thus increased availability of memory tends to favor compiled strategies, whereas increased processing capability favors interpretive strategies. As has been suggested, both will be needed for high performance systems of the future. It is well to remember that technology is progressing much faster than knowledge of speech. For this reason, a *correct theory* is the central need of research. The technology will easily rise to support such a theory. Through comprehensive study of extensive databases, new and insightful complex abstract unit hierarchies will be developed which when interpreted procedurally within domains of varying scope can be expected to give rise to synthetic speech of very high quality.

tence, clause, phrase, word, morpheme, metrical foot, syllable, phoneme, and feature. These have all been found useful in linguistic analysis for reasons of distribution and contrast, but they each exhibit some intrinsic cohesion and place focus on the units themselves. Thus each of these units implies an internal theory, together with a relation to other units and their corresponding theory. There is a tendency to think of these different levels of units with their attendant scopes as being arrayed over a temporal axis. While such a view provides a convenient geometric intuition, contemporary notions of abstraction may not rely on notions of temporal scope at all levels. Thus at the individual segment feature level the initial specification may include an unordered set of features which is then transformed at a higher level to derive temporal extent. Research in the asynchronous control of these features reveals the need for this varying scope at the feature level, as in the control of nasality.

11. The hierarchy of units should provide for a *gradual* progression through the various representations. That is, a sufficient number of units should be available so that the transformation from one level to another through the hierarchy is relatively small and direct. Thus for example, phonemes are transformed to allophones and then to targets, parameters and finally to the speech waveform. In this way, understanding of the speech process is modularized at each level, thus simplifying the rule structure at each level, even though the total number of rules over the entire system may be large.
12. The notion of *target* serves as a bridge between abstraction and physical properties. There is a tendency for the targets themselves to become complex objects exhibiting composition, scope, and internal cohesion. For example, fundamental frequency contours must be characterized at both the phrase level and in terms of 'segmental' effects. It is seen that the term 'target', originally motivated by the characterization of formant trajectories, is increasingly synonymous with 'abstract representational structure'. The role of *variability* in the choice of units must be considered in the design of any speech synthesis system. It is often felt that many attributes of these units are necessary, and hence admit of no variability. Recent understanding, however, indicates that many cues are redundant, and that some contextually determined integrative process that may be deterministic in nature determines the cues that must be instantiated in order to derive the intended percept. Thus there may be a level of variation on the surface that appears to be free or random, but which in fact is a highly complex manifestation of underlying regularity. There are no contemporary speech synthesis systems that begin to approach the level of surface phonetic variability observed in natural speech. Models are needed that characterize this variability in terms of redundant cues, idiosyncratic gestures associated with a given speaker, and true random variability. The level of understanding needed to provide this control is immense, and will doubtlessly elude practical systems for many years to come.

13. The units of speech synthesis should be able to support a wide class of *voices*. Most contemporary systems focus on the adult male voice, but child, and female voices must be obtainable by any general system, and there is a need for the ability to synthesize several languages and diverse dialects within the context of one overall synthesis framework. A few contemporary systems are capable of synthesizing utterances in several languages within the same framework, but these systems generally sacrifice quality of synthesis in any particular language for breadth of coverage over the diverse linguistic forms.
14. The choice of units is intimately related to questions of *intelligibility*, *naturalness*, and *reduction* phenomena. To date, most research in speech synthesis has aimed at producing intelligible speech, without as much attention being focused on issues of naturalness and vowel reduction. There is still much to be learned about vowel reduction, and some feeling that such reduction may come from redundancy when viewed in the context of the sum total of all constraints in force at any point in the utterance. There is a tension between intelligibility and naturalness in that if a system aims for substantial reduction in the name of naturalness, intelligibility may suffer. It is also very difficult to evaluate naturalness, although some success has been achieved using cognitive loading techniques. Nevertheless, this class of psycholinguistic experimentation is still in its infancy and a great deal of work needs to be performed. At the discourse level, there is increased need for naturalness since attributes of focus and old/new information cause substantial stressing and destressing over that which would be predicted for a sentence spoken in isolation. It should also be remarked that individual segment intelligibility still leaves much to be desired, and the codification within an insightful rule structure of the vast amount of segmental phonetic detail must be substantially improved if there is to be any appreciable gain in synthetic speech quality.
15. The units may exploit available technology, but they should not be overly determined by it. Computational technology typically involves space/time tradeoffs. Thus increased availability of memory tends to favor compiled strategies, whereas increased processing capability favors interpretive strategies. As has been suggested, both will be needed for high performance systems of the future. It is well to remember that technology is progressing much faster than knowledge of speech. For this reason, a *correct theory* is the central need of research. The technology will easily rise to support such a theory. Through comprehensive study of extensive databases, new and insightful complex abstract unit hierarchies will be developed which when interpreted procedurally within domains of varying scope can be expected to give rise to synthetic speech of very high quality.