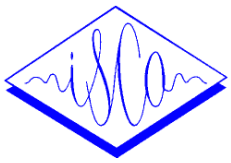

The ESCA Workshop
on Speech Synthesis
Autrans, France
September 25-28, 1990

DERIVING TEXT-TO-SPEECH DURATIONS FROM NATURAL SPEECH.

Jan P. H. van Santen

Linguistics Research Department
AT&T Bell Laboratories
Murray Hill, New Jersey 07974, USA

ABSTRACT

Any text-to-speech system has a subsystem (*duration system*) that computes speech timing. How does one construct a duration system that accurately mimics natural speech? This paper discusses a particular type of data analysis method for the statistical analysis of natural speech durations, *ordinal data analysis*, and shows how it can be used for the construction of duration systems.

1. INTRODUCTION

Text-to-speech synthesis systems require a *duration system*, a module that computes the duration of any segment in any prosodic context. Currently, at least four types are used. The best known is the *sequential rule* system where rules of the type

If vowel v is stressed:
output duration = T(input duration; stressed,v)

are applied successively, starting with a base duration [Klatt, 1987]. In Klatt's system $T()$ has the form $\alpha[stressed] \times (input\ duration - min[v]) + min[v]$, where $\alpha[stressed]$ is a multiplier and *min[v]* is the minimum duration of vowel *v*. A second approach is based on *lookup tables* that list durations for each of all possible context / segmental identity combinations (I call these combinations the *cells* of the *factorial space* generated by the contextual factors and the segmental identity factor). An example of a lookup table entry would be

Stressed /i/ in final syllable of accented phrase-medial polysyllabic word preceded by a voiced stop followed by /m/ and followed by a clitic: 185 ms.

A third method uses *equations* to mathematically combine scale values for segmental and contextual factors [e.g., Coker, Umeda, & Browman, 1973]. For example, in the equation

$$d(x,y_1,y_2,\ldots,) = g(x) + h_1(y_1) + h_2(y_2) + \cdots \quad (1)$$

$d()$ denotes duration, $g(x)$ the scale value of segment x, and $h_i(y_i)$ the scale value for level y_i on factor Y_i. E.g., if Y_2 is *voicedness of the postvocalic consonant*, then the levels are {*voiced*, *unvoiced*} and the corresponding scale values $h_2(voiced)$ and $h_2(unvoiced)$ might be 50 ms and 10 ms, thus generating a 40 ms voicing effect. A fourth type of system uses *binary trees*, with durations as "leaves" [Riley, 1989]. Here, nodes correspond to dichotomies such as "postvocalic consonant is fricative vs. non-fricative"

Although the differences between these systems appear profound, there are important equivalencies. First, any system can generate a complete lookup table by tabulating its computed durations for all cells. Second, any system can in principle be written as an equation, although the equation can be prohibitively cumbersome for lookup table systems and tree-based systems; nevertheless, it is important to realize that any system has an *implicit equation*. For example, for the Klatt [1987] rule system the implicit equation has the form

$$d(x,y_1,y_2,\ldots,) = \alpha_1[y_1] \times \alpha_2[y_2] \times \cdots \times (inh[x] - min[x]) + min[x], \qquad (2)$$

where *inh[x]* is called the *inherent duration of x.*

How do we construct a duration system that generates durations that are as close as possible to natural durations? I propose here that, while ultimately the quality of a duration system is a perceptual matter, a critical first step consists of the statistical analysis of natural speech durations. This paper focuses on a particular type of data analysis method, called *ordinal data analysis*, and discusses how this method can be used for the construction of duration systems.

2. ORDINAL DATA ANALYSIS

Segmental duration in natural speech is notoriously complex, because the duration of a segment depends on many factors and because the effect of a factor, whether measured in milliseconds or as a percentage change, typically depends on other factors: Factors interact. These interactions are generally not chaotic, however, since often one factor leaves the direction of the effects of another factor unaltered -- only amplifies its effects. To illustrate, the effects of the postvocalic consonant class (defined in terms of manner of production and voicing) on stressed vowel duration are proportionally much larger in phrase-final than in phrase-medial position, but in both positions the largest vowel durations are obtained for voiced fricatives, next for voiced stops, and, at the short end, for unvoiced fricatives, and, finally, for unvoiced stops and affricates. Factors whose direction of effects remains the same are called *monotone* [Krantz, Luce, Suppes, and Tversky, 1971]

Because most duration factors tend to be monotone (e.g., vowel identity, syllabic stress, word length, phrasal boundary, speaking rate), exceptions are especially informative. For example, dental consonants tend to be shorter than velar consonants at the head of unstressed syllables but longer at the head of stressed syllables [Umeda, 1977]. Although this non-monotonicity obviously is due to flapping of dental consonants, there exist less obvious cases of non-monotonicity that are not immediately understood and hence may yield new insights in the phenomena. An additional benefit from inspection of monotonicity is that it may lead to a re-thinking of the factors, such as giving flapped dentals separate symbols on the segmental identity factor.

Now, the property of monotonicity is an example of an *ordinal pattern*, i.e., a pattern that has to do with the way durations are ordered across the cells. Inspecting whether a factor is monotone is an example of *ordinal data analysis*. There exist statistical methods to determine whether violations of monotonicity are systematic or are due to "noise" in the data [van Santen, 1989; van Santen & Olive, 1989, 1990].

Another example of an ordinal pattern is *joint independence* [Krantz et al., 1971]. Suppose

$$d(/i/,voiced,m) > d(/i/,unvoiced,m) > d(/I/,voiced,m) > d(/I/,unvoiced,m)$$

but

$$d(/i/,voiced,f) > d(/I/,voiced,f) > d(/i/,unvoiced,f) > d(/I/,unvoiced,f),$$

where *f* (*m*) denotes phrase-final (phrase-medial) position and where voicing refers to the postvocalic consonant. Here, voicing and vowel identity are monotone, but the direction of their joint effects depends on position. This is an example of a violation of joint independence. On the other hand, syllabic stress and vowel identity tend to produce the same joint order in phrase-final and phrase-medial position, and thus are jointly independent.

As with monotonicity, inspecting for which factors joint independence holds can cast light on underlying processes. For example, the fact that phrase boundaries amplify the effects of postvocalic voicing more than the effects of segmental identity and stress may be related to phrase boundaries and postvocalic voicing primarily affecting the final portion of vowel segments and lexical stress affecting other portions as well.

The above illustrates that ordinal patterns can be of substantive importance and that duration data, far from being chaotic, have quite a bit of ordinal structure. An additional fact about ordinal patterns is that they tend to be more invariant than quantitative aspects of the data. That is, data sets may differ in overall speaking rate, in the magnitude of the effects of certain factors, and in the sizes of various interactions; however, ordinal patterns tend to be relatively invariant. For example, the violation of joint independence of vowel identity and postvocalic voicing across phrasal positions can be found in many data sets, despite large differences in other respects.

3. ORDINAL DATA ANALYSIS AND DURATION SYSTEMS

How can ordinal data analysis be used to construct a duration system? The key lies in the implicit equation of a duration system [e.g., Eq. (2)]: *Ordinal patterns can be used to choose among very large classes of equations* [van Santen, 1989; van Santen & Olive, 1989, 1990]. Let me illustrate this by providing an elementary proof that monotonicity is necessary for the additive model [Eq. (1)] to hold. For the segmental identity factor we have:

$$
\begin{aligned}
d\,(/i/,y_1,y_2) &\geq d\,(/I/,y_1,y_2) && \textit{if and only if}\\
g\,(/i/) + h_1(y_1) + h_2(y_2) &\geq g\,(/I/) + h_1(y_1) + h_2(y_2) && \textit{if and only if}\\
g\,(/i/) &\geq g\,(/I/) && \textit{if and only if}\\
g\,(/i/) + h_1(y'_1) + h_2(y'_2) &\geq g\,(/I/) + h_1(y'_1) + h_2(y'_2) && \textit{if and only if}\\
d\,(/i/,y'_1,y'_2) &\geq d\,(/I/,y'_1,y'_2).
\end{aligned}
$$

The same proof can be given for the other factors. To illustrate what is meant by "very large", it can be shown that monotonicity is necessary not only for the additive model but for any equation of the form

$$d\,(x,y_1,y_2,\ldots,) = F\,[g\,(x) + h_1(y_1) + h_2(y_2) + \cdots], \qquad (3)$$

where $F\,[]$ is an arbitrary increasing function (for the additive model, $F\,[y\,] = y$). This covers a great variety of equations, including also the multiplicative model (where $F\,[y\,] = e^y$). In fact, monotonicity of one particular factor is implied by even larger classes of equations. For example, the implicit equation of the Klatt duration system [Eq. (2)], which is not of the form shown in Eq. (3), implies monotonicity of all factors with the exception of the segmental identity factor.

It can also be shown that violations of joint independence contradict the additive model, the multiplicative model and a host of other models. (Thus, a violation of joint independence can be seen as more general type of interaction than the usual additive interaction in the Analysis of Variance.) Conversely, when a set of factors is jointly independent, then this also restricts which equations apply; in particular, if all possible forms of joint independence hold then, if certain additional conditions are met, only equations of the form given in Eq. (3) apply [Krantz et al, 1971].

How do we construct a duration system whose implicit equation is compatible with observed

ordinal patterns? For the equation approach, the procedure consists of formulating a family of equations compatible with the observed ordinal patterns, and selecting the best-fitting member of this family with model-fitting methods [van Santen & Olive, 1990].

For the lookup table approach, one has to use a branch of statistics called *estimation under order restrictions* [Barlow, Bartholomew, Bremner, and Brunk, 1972]. These methods make it possible to create a lookup table that satisfies a given set of ordinal patterns, even when there are many cells with few or no observations.

For sequential rule systems one has to construct a preliminary rule system, compute its implicit equation, and verify if this equation is compatible with the ordinal data patterns. At times it can be quite clear what changes must be made. For example, if all rules pertaining to factors X precede all rules pertaining to the remaining factors, then the X factors should be jointly independent; if the data tell otherwise, a simple change in the order of the rules may be sufficient. On other occasions, in particular with complicated rule systems, it may be less clear what changes to make.

The most difficult problem for the incorporation of ordinal patterns is posed by tree-based systems. Although, as with any other system, one could use estimation under order restrictions to rectify the lookup table that this system generates, a more satisfactory but unexplored approach is to build the ordinal patterns into the tree growing process itself.

4. FINAL REMARKS

The key advantage of ordinal data analysis over other data analysis methods is that it allows drawing precise inferences that are not tied to a particular narrowly defined family of equations. In a way, ordinal patterns *reflect the fundamental structure of the interactions in the data.* To the extent that ordinal patterns are perceptually significant, it is important to use duration systems that can incorporate these patterns.

REFERENCES

Barlow, R.E., Bartholomew, D.J., Bremner, J.M., & Brunk, H.D. (1972). *Statistical Inference under Order Restrictions.* New York: Wiley.

Coker, C.H., Umeda, N., & Browman, C.P. (1973), "Automatic synthesis from ordinary English text". IEEE Transactions on Audio and Electroacoustics, AU-21, 3, 293-298.

Klatt, D.H. (1973), "Interaction between two factors that influence vowel duration". J. Acoust. Soc. Am., 54(4), 1102-1104.

Klatt, D.H. (1987), "Review of text-to-speech conversion for English", J. Acoust. Soc. Am., 82(3), 737-793

Krantz, D.H., Luce, R.D., Suppes, P., & Tversky, A. (1971), "Foundations of Measurement", Vol. I, New York: Wiley.

Riley, M.D. (1989), "Statistical tree-based modeling of phonetic segment durations", J. Acoust. Soc. Am., 85, S44 (Q8).

van Santen, J.P.H. (1989), "Two diagnostic tests for vowel duration models", Paper presented at the 22nd Annual Mathematical Psychology Meeting (Irvine, CA, August 3-5 1989).

van Santen, J.P.H., & Olive, J. (1989) "Diagnostic tests of segmental duration models", J. Acoust. Soc. Am., 85, S43 (Q1).

van Santen, J.P.H., & Olive, J. (1990) "The Analysis of Contextual Effects on Segmental Duration", Computer Speech & Language (in press).