EUROSPEECH '97
5th European Conference on Speech Communication and Technology
Rhodes, Greece, September 22-25, 1997

Analysis of Speaking Rate Variations in Stress-timed Languages

*Tom Brøndsted and Jens Printz Madsen**
Center for PersonKommunikation
Aalborg University, Fredrik Bajers Vej 7 A2, DK-9220 Aalborg Øst, Denmark.
Tel. +45 96 35 86 36, FAX: +45 98 15 15 83, E-mail: tb,@cpk.auc.dk.

Abstract

This paper analyses speaking rate variations in English and Danish and relates them to problems encountered in speech recognition. Intra speaker variabilities in speech rates are explained with reference to time equalisation of stress groups and utterances. Further, it is shown that certain natural classes of phonemes are more affected by speaking rate variations than others.

Keywords: Phoneme modeling, rate of speech, phone duration, time equalisation of stress groups and phrases.

10.21437/Eurospeech.1997-163

1. Introduction

Large vocabulary Speech recognition systems based on Hidden Markov Models modeling phonemes or units derived from phonemes (triphones, generalised triphones, diphones) have over the recent years moved towards increasing feature vector dimensions which typically include parameters like 1st and 2nd order delta cepstrum. Furthermore preprocessing is typically performed within a fixed window of around 100 msec necessary for modeling the huge number of qualitative phoneme variants found in large training databases. However, the number of acoustic events within such a window or speech segment is dependent on ROS (Rate of Speech). Since preprocessing includes speed dependent parameters, the acoustic model itself becomes ROS dependent.

Figure 1 shows the typical correlation between ROS measured simply as phones/second, henceforth ROS_{PH}, and word accuracy obtained with phoneme-based recognition of a Danish database P1 (Brøndsted 1994):

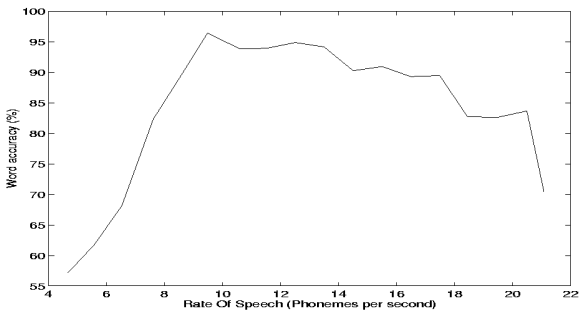

Figure 1: ROS_{PH} versus Word Accuracy in P1

Performance decreases in case of very slow or very fast speech, whereas the best results are obtained on "normal" speech rates at about 10-17 phones/sec. This problem has been addressed in papers like Macchi et al 1990, Mirghafori et al 1996, however, we feel that it calls for a more profound analysis. ROS_{PH} depends not only on speaker specific habits ("fast" vs. "slow" speakers) but also on inherent properties of the utterance spoken. Further, not all phoneme durations are equally affected by ROS_{PH} variations. The present paper explores these two issues. Our analysis is carried out on the American TIMIT database (Gorofolo et al. 1993) and a small vocabulary Danish database P1 (Brøndsted et al. 1994)

2. Stress Groups, Utterance Lengths vs. ROS_{PH}

ROS_{PH} is not a very speaker specific measure. Our analysis shows a relatively high intra-speaker variability of ROS_{PH} in both TIMIT and P1, though both databases were recorded under constant external conditions (this means that we can leave situative factors as varying degrees of eagerness, anger etc. out of account). Consequently, ROS_{PH} must be dependent also on certain inherent properties of utterances, like (1) the actual phonological constituents of the utterance (as, for instance, phonologically long vowels and diphthongs by nature are longer than short and unstressed ones) and (2) the length of stress groups and the length of the entire utterance. The present paper concentrates on aspect (2).

In general phonetics, it is assumed that stress-timed languages like English and Danish (as opposed to syllable-timed languages like French) tend to have a relatively constant duration of stress groups, independent of the actual number of phones or syllables involved in these groups (cf. Grønnum 1992). Consequently, we may expect the time duration between the capitalized syllables in e.g. (a) "the BUS from LEEDS" and (b) "the BUSes from the NORTH" to be approximately the same when spoken by the same speaker under the same external conditions. In terms of ROS_{PH}, (a) would be a "slow" phrase (few phones and syllables per stress group), whereas (b) would be "fast" (many phones and syllables per stress group). Further, in terms of stress groups per second, henceforth ROS_{SG}, we may not expect any significant difference between (a) and (b). In stress-timed

* In alphabetic order.

languages ROS_{SG} may be a more speaker specific measure than ROS_{PH}.

To test this theory, the utterances spoken by two speakers in the Danish P1 database were segmented into stress groups defined simply as a group of syllables, the first of which has the primary stress and the subsequent syllables are secondary stressed or unstressed. The two speakers chosen for the analysis where the one with the fastest and the one with the slowest ROS_{PH}. However, the result shown in figure 2 f., does not fully confirm the theory.

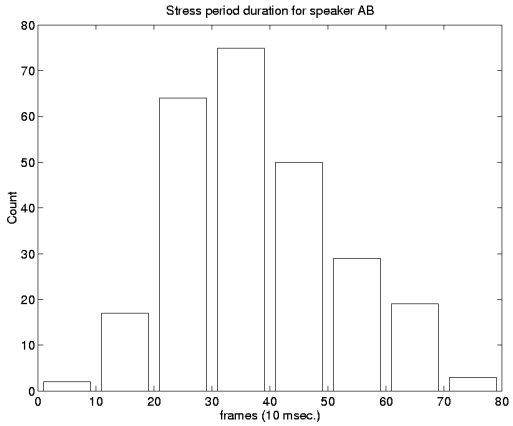

Figure 2: Stress group duration for fast speaker in P1.

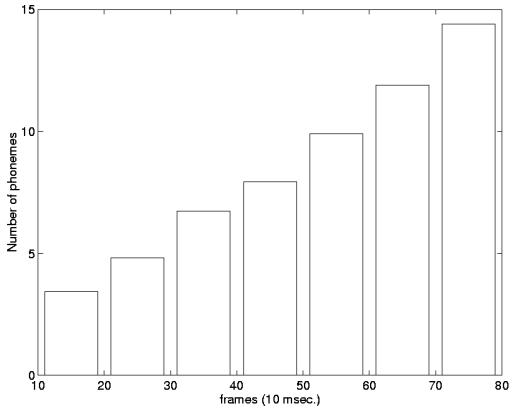

Figure 3: Stress group duration (msec.) vs. number of phonemic constituents in P1.

Not only does the duration vary considerably for the same speaker - between 300-700 msec. for the "slow" and 200-600 for the "fast" speaker (fig. 2) -, but there is further a rather linear correlation between stress group durations and the actual number of phonemes constituting the groups (fig. 3). We ascribe this to the fact, that P1 (as also TIMIT) contains read speech and that the P1 sentences were generated randomly from a predefined grammar (APSG) and in many cases have abnormal semantics. In spontaneous speech, we may expect the correlation between stress group duration and length in terms of number of phonemes to be less linear.

However, in both TIMIT and P1, we found a significant correlation between ROS_{PH} and the length of the entire spoken utterance measured as the total number of phonemic constituents. Figure 4 shows that long sentences are spoken faster than short ones:

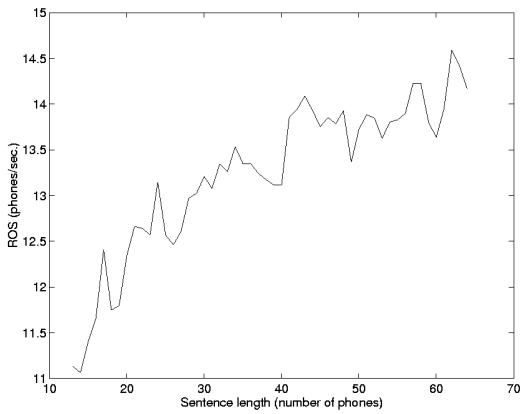

Figure 4: Sentence length (number of phonemes) vs ROS_{PH} in TIMIT

The tendency can be explained by "the law of equalisation", i.e. the speaker's endeavor to pronounce short and long utterances in approximately equal time (cf. Malécot et al. 1972, Fónagy et al. 1960). In P1, and probably also in TIMIT, intra-speaker variabilities of ROS_{PH} are more due to variations of sentence lengths than to that of stress group lengths.

3. Phoneme Duration and ROS_{PH}

To estimate the dependence of an individual phoneme on ROS_{PH}, we start from two measures. The first one $ROS_S(r)$ describes the rate of speech of an individual sentence $S(r)$ as the average duration of its phoneme manifestations (phones):

$$ROS_S(r) = \frac{1}{N_S(r)} \sum_{i}^{S(r)} \frac{1}{dur(r,i)} \quad \textbf{(1)}$$

where $N_S(r)$ is the number of phones constituting the sentence $S(r)$ and $dur(r,i)$ is the duration of phone number i. This measure largely corresponds to the ROS definition suggested by Mirghafori et al. (1996). The second measure $ROS_P(r,l,j)$ describes the speech rate of an individual phone (j) in the sentence $S(r)$, i.e.

$$ROS_p(r,l,j) = \frac{1}{dur(r,l,j)} \quad \textbf{(2)}$$

where $dur(r,l,j)$ is the duration of the phone segment number l in the sentence $S(r)$ transcribed by the phoneme symbol (j). The actual dependence of a phoneme on ROS_{PH}-variations is calculated via first order regression coefficients:

$$y = f_j(x) = a_j + b_j x \quad \textbf{(3)}$$

which are estimated on the data sets:

$$(x, y)_j : (ROS_S(r), ROS_p(r, l, j)) \quad \textbf{(4)}$$

where $(x,y)_j$ is the data set for each phonemic symbol $p(j)$ and where the regression is performed on each of these sets. The dependence measure can now be defined as the relative change in $ROS_P(r,l,j)$ in respect to $ROS_S(r)$, and given by:

$$R(j) = \frac{\frac{f_j(x_0 + \Delta_x) - f_j(x_0)}{f(x_0)}}{\frac{(x_0 + \Delta_x) - x_0}{x_0}} = \frac{x_0}{y_0} b_j \quad \textbf{(5)}$$

where (x_0,y_0) is the data point which the measure $R(j)$ will be based on. The data sets in TIMIT for the phonemes /b/ and /uw/ (IPA: the release phase of /b/ and the slightly rising diphthong /u^w/) are shown in figure 5 below:

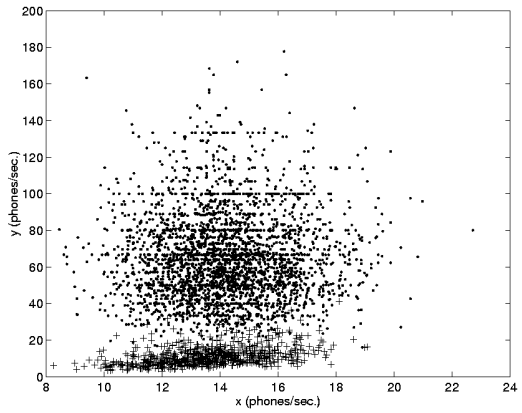

Figure 5: ROS data sets in TIMIT for the phonemes /b/ and /uw/ marked with (.) and (+), respectively.

The /b/ manifestations distribute themselves over the entire cluster demonstrating no significant sensitivity to ROS_{PH} variations ($R(j)$=0.04), whereas the manifestations of /uw/ are distributed close to the x-axis with a rising tendency ($R(j)$=1.49), indicating a distinct ROS_{PH} sensitivity[1] .

The measure $R(j)$ presupposes large data sets. Consequently, we only give the results from the analysis of TIMIT. For each phonemic symbol, we further calculate a statistical confidence score $C(j)$, which roughly expresses the extent to which we can rely on the calculated $R(j)$ value[2] . In the table below, each phoneme $P(j)$ is entered with the original TIMIT-symbol (a simple ASCII-transcription of IPA). A bracket () indicates that the symbol denotes a "deviant" pronunciation of a phoneme (for details, see Brøndsted 97). Finally, each phoneme's association with a natural class of segments is indicated by 'X': *gl=glide, lv=long vowel or diphthong, liq=liquid, cl=closure phase (i.e. of affricates or plosives), sv=short vowel, fr=fricative, na=nasal, aff=affricate (delayed release phase), pl=plosive (release phase)*[3] .

P(j)	R(j)	C(j)	gl	lv	liq	cl	sv	fr	na	af	pl
uw	1.49	0.62		X							
w	1.47	0.25	X								
(ux)	1.35	0.34		X							
y	1.33	0.35	X								
oy	1.21	0.57		X							
r	1.14	0.18			X						
ao	1.02	0.25		X							
th	1.01	0.50						X			
axr	0.99	0.24					X				
ae	0.96	0.23					X				
ow	0.95	0.30		X							
ng	0.92	0.37							X		
bcl	0.92	0.32				X					
en	0.90	0.52							X		
(hv)	0.89	0.43						X			
s	0.89	0.17						X			
tcl	0.89	0.18				X					
hh	0.88	0.43						X			
aw	0.88	0.49		X							
uh	0.87	0.52					X				
dcl	0.87	0.21				X					
aa	0.84	0.24					X				
ay	0.82	0.29		X							
l	0.82	0.17			X						
el	0.80	0.44			X						
ih	0.79	0.19					X				
(q)	0.77	0.25									X
kcl	0.75	0.19				X					
zh	0.73	0.94						X			
ey	0.72	0.27		X							

[1] The clear horizontal lines in the distribution of /b/ in figure 5 could indicate that plosives have been segmented automatically in closure and release and subsequently manually adjusted if necessary.

[2] The confidence score $C(j)$ is given by

$$C(j) = \frac{x_0}{y_0} \sqrt{\frac{1}{S_{xx}(j)}} \sqrt{\frac{SS_R(j)}{N_j - 2}} t_{\frac{\alpha}{2}, N_j - 2}$$

where Nj is the number of occurrences of the phoneme $p(j)$ in the database. SS_R and S_{xx} are defined as in Ross Sheldon, 1987.

[3] Not all vowel symbols are unequivocal with respect to the phonological feature *tense*. In particular /ER/ which in TIMIT is used in both e.g. *backwards* and *birds* (/b ae1 k w er d z/, /b er1 d z/) can be classified as both a short and a long vowel. For details, see Brøndsted 1997.

er	0.72	0.30		X							
g	0.72	0.32									X
n	0.70	0.17							X		
f	0.67	0.30						X			
z	0.66	0.22						X			
iy	0.66	0.16		X							
pcl	0.66	0.27				X					
eh	0.66	0.22					X				
(eng)	0.64	1.70							X		
v	0.64	0.31						X			
ah	0.63	0.27					X				
ix	0.63	0.15					X				
ch	0.63	0.47								X	
ax	0.57	0.22					X				
m	0.54	0.23							X		
(axh)	0.51	0.74					X				
gcl	0.50	0.31				X					
sh	0.48	0.30						X			
jh	0.48	0.40								X	
(em)	0.47	1.16							X		
dh	0.47	0.28						X			
t	0.38	0.23									X
p	0.21	0.29									X
(nx)	0.20	0.44							X		
(dx)	0.17	0.26									X
d	0.17	0.24									X
k	0.12	0.22									X
b	0.04	0.28									X

The horizontal order of natural classes as well as the vertical order of individual phonemes are defined by the *R(j)* tendency. Consequenly, the 'X's largely form a diagonal line from the upper left to the lower right corner of the table. The theory of naturalness (as defined by Chomsky & Halle 1968) allows of using alternative phonological classifications (e.g. various cavity and source features), however apparently the clearest relation between naturalness and *R(j)* is given by the suggested classification into traditional "major classes". The natural classes used in the table above are defined by the major class distinctive features *sonorant, syllabic, consonantal* together with the manner of articulation features *tense* and *continuant*. Phonological features like *high, back, low, anterior, coronal, voice* etc. are apparently of no importance to *R(j)*.

As mentioned above, the Danish P1 database is too small to allow a similar analysis of *R(j)*, however the tendency seems to be the same as in TIMIT. For details, see Brøndsted & Madsen 1997.

4. Conclusion

The analysis of the P1 and TIMIT databases shows that ROS_{PH} is not a adequate measure for speech rate, as we from an adequate measure must expect a low intra-speaker variability in speech databases recorded under constant external conditions. The theory of time equalisation points in the direction of other and longer units of ROS than phones: In stress-timed languages like Danish and English: stress groups, phrases, and entire utterances. In TIMIT we have shown the effects of time equalisation of entire utterances. Our analysis has failed to show the same effect on stress groups. We ascribe this to the fact that the analysed databases contain read speech.

For the analysis of the problems of ROS variations encountered in speech recognisers modeling phonemes or units derived from phonemes (diphones, triphones), ROS_{PH} is a useful measure. The problem of varying lengths of phoneme manifestations is not evenly distributed over all phonemes. We have suggested and described a measure to determine a phoneme's dependence on ROS_{PH} variations. Most affected by such variations are glides and phonological long vowels. Almost unaffected are plosives and affricates.

The authors wish to thank Jürgen Trouvain, Universität des Saarlandes, for valuable discussions.

5. References

Brøndsted, T. Larsen, L.B.: Representation of Acoustic and Linguistic Knowledge in Continuous Speech Recognition. Technical Report R 94-1001. CPK, Aalborg. Spoken Language Dialogue Systems 1994.

Brøndsted, T.: A SPE based Distinctive Feature Composition of the CMU Label Set in the TIMIT Database. Internal draft, Aalborg 1997.

Brøndsted, T., Madsen, J.P.: Fonemteori og talegenkendelse. In Brøndsted & Lytje (eds.): Sprog og multimedier. Aalborg Universitetsforlag 1997.

Chomsky, N., Halle, M.: The Sound Pattern of English. 1968.

Fónagy, I., Magdics, K.: Speed of Utterance in Phrases of different Lengths. Language and Speech 1960.

Garofolo, J.S. et al.: DARPA TIMIT. Acoustic-Phonetic Continuous Speech Corpus. NistIR 4930 1993.

Grønnum, N.: The Groundwork of Danish Intonation. Museum Musculanum Press, Copenhagen.

Macchi, M.J., Spiegel, M.F., Wallace, K.L.: Modeling Duration Adjustment with Dynamic Time Warping. IEEE Int. Conf. Acoustic Speech, Signal Processing, 1990.

Malécot, A., Johnston, R., Kizziar, P.A.: Syllabic Rate and Utterance Length in French. Phonetica 26, 1972.

Mirghafori, N., Fosler, E., Morgan, N.: Towards Robustness to Fast Speech in ASR. IEEE Int. Conf. Acoustic Speech, Signal Processing, 1996.

Ross Sheldon, M: Introduction to Probability and Statistics. John Wiley and Sons, Inc., 1987.