Articulation Rate Measures and Their Relation to Phone Classification in Spontaneous and Read German Speech

Jürgen Trouvain, Jacques Koreman, Attilio Erriquez & Bettina Braun

Institute of Phonetics, University of the Saarland, Saarbrücken, Germany
{trouvain,koreman,erriquez,bebr}@coli.uni-sb.de

Abstract

This paper evaluates articulation rate measures and rate characteristics of read and spontaneous speech on the basis of a manually labelled database for German. The results of phone classification experiments for three different articulation rates only partially confirm our expectations. Phonetic explanations are suggested.

1. Introduction

It has been observed in several studies that speech rate strongly affects the recognition rates of automatic speech recognition (ASR) systems. This is particularly true for fast speaking rates, as has been shown for several languages [1–5], as well as for slow, hyper-articulated speech [1,3,6]. Figure 1 illustrates the relationship between word error rate and articulation rate in a stylised picture.

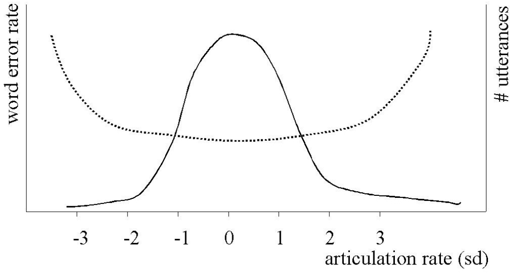

Fig. 1: Stylised correlation between articulation rate and recognition rate and number of utterances, cf. [1].

The aim of this paper is to evaluate the relationship between articulation rate and phone classification for German read and spontaneous speech. Our analysis uses *linguistic* units derived from a labelled database. We explicitly do not address the problem of on-line estimation of acoustic measures for speech rate, see e.g. [7].

The rest of this paper is divided into three parts. Section 2 discusses measures of articulation rate derived from a labelled database. In section 3 we describe the articulation rate characteristics of read and spontaneous speech, both in terms of measured articulation rates and in terms of the phonemic variation underlying these measures. In section 4 we discuss phone classification rates for three different articulation rates.

2. Quantification of Articulation Rate

A prerequisite for any approach to rate modelling in ASR is a classification of the speech data into its articulation rate characteristics. In this study we shall compute several speech rate measures and evaluate them with regard to temporal variance, comparing German read and spontaneous speech.

2.1. Database

The German KielCorpus for Read and Spontaneous Speech [8] contains *manually* labelled realised phones along with information on the lexical status for each phone. The read speech data (4 hours; 53 speakers) consists mostly of single sentences of varying length and two short stories. The spontaneous speech data (4 hours; 52 speakers) consists of recordings of appointment-making dialogue. The larger part of the database was also labelled for prosodic phrase boundaries as well as for pitch accents and contours. For this study the parts labelled segmentally *and* prosodically were selected.

2.2. Articulation rate measurements

Two questions are addressed here. The first question is which linguistic unit best reflects temporal variation. The second question is which domain is appropriate. The measure we use for this is the strength of the correlation between duration of the domain and the number of units in the domain.

2.2.1. Linguistic units

A number of different linguistic units are in use for measuring speech rate such as the word, the syllable, and the phone. Although commonly used, the definition of these linguistic units is not always straightforward. In particular, although *intended* forms have the advantage that they can be easily derived from the lexical representation of the uttered words, their actual *realisation* can vary strongly. The German sentence "Am blauen Himmel ziehen die Wolken." (Engl. lit. "In the blue sky wander the clouds.") consists of the following transcription of 26 phonemes and 10 syllables: /?am blaU@n hIm@l tsi:@n di: vOlk@n/[1]. However, a typical reduced realisation of this sentence in the corpus, [am blaUn hIml tsini vOlkN], is shrunk to 20 phones in 8 syllables.

Not only is there a difference between intended forms and their actual realisations, there is also disagreement about the definition of what is and is not a linguistic unit. On the phone level, glottal stop is considered as a phone in the database, affricates have been counted as two sounds, diphthongs have been regarded as one sound. Vowel-/r/-combinations have been counted as two sounds in the intended form and only one sound in the diphthongised realised form. Schwa-/r/-combinations have been labelled as one phone /6/ in both representations. Counting the number of intended syllables was rather unproblematic, whereas the decision whether a syllable is realised or not is not that easy (e.g. the /@n/ syllable in “ziehen” in the example above can be realised as a syllabic or

[1] The SAM Phonetic Alphabet is used here. [9]

non-syllabic /n/, leading to different syllable counts). On the word level we opted for the graphical word separated by a space and/or punctuation marks, although e.g. the morphological word would also be a possible unit.

For this study we counted the following units: intended words, intended syllables, realised syllables, intended phones and realised phones.

2.2.2. Domain

Another uncertainty when dealing with speech rate concerns the stretch of speech taken into consideration. Speech rate changes continuously while speaking [10], so that the first part of an utterance can be spoken fast, while the second part can be rather slow. An average rate calculated for an utterance does not reflect the tempo characteristics of different parts. When the domain is not specified, it is not clear whether the speech rate quantifications are related to a more global or to a more local level.

Since ASR is primarily interested in decoding speech from the information contained in its phone segments, pauses (and their potential informational content) are often ignored. Pauses are a very important mechanism for the speaker to vary speech tempo [11].

Pauses normally occur at prosodic-syntactic boundaries in read speech, whereas in spontaneous speech additional hesitation pauses represent breaks of performance units. They are easy to determine and therefore often used to delimit the domain over which *articulation* rate is calculated. Articulation rate *ex*cludes pauses, distinguishing it from *speaking* rate, which *in*cludes pauses.

In this study the linguistic units were computed for the following two utterance domains: for the above mentioned reasons the *inter-pause stretch* (ips) was selected as the articulation phase between two pauses (including silence, breathing, filled pauses and other non-verbal articulations like coughing and lip smacking). Like the ips, the *intonation phrase* IP (with only one level of phrase boundary strength) is considered as an important planning unit, reflected by the intonation contour. Note that in spontaneous speech hesitation pauses can interrupt intonation phrases. Thus, an IP can consist of multiple ips, just as an ips can consist of multiple IPs.

2.3. Correlation analysis

For an optimal articulation rate measure the number of linguistic units in the domain should correlate highly with the duration of the domain. The results presented in table 1 show high and significant correlations for all selected units and for both utterance domains, both for read and for spontaneous speech.

Table 1: Correlations of number of linguistic unit with articulation time, broken down for each linguistic unit and domain, for spontaneous and read speech. All correlations are significant at p<0.001.

	spontaneous		read	
unit	ips	IP	ips	IP
intended word	.913	**.855**	.899	**.862**
intended syllable	.951	.926	.938	.918
realised syllable	.955	.932	.945	.924
intended phone	.956	.935	.945	.928
realised phone	**.965**	.948	**.958**	.943

The observed correlations are systematically higher for the ips than for the IP. In spontaneous speech, the 'realised phone' shows the highest correlation with duration (r=0.965 for ips) while the lowest correlation is found for the 'word' (r=0.855 for IP). The same pattern is found in read speech.

2.4. Discussion

Despite the high correlation of the 'realised phone' with duration, other criteria for an *optimal* linguistic quantification of speech rate can be applied. The 'word' and the 'intended syllable' have the advantage that they only need an orthographic transliteration, whereas the 'realised phone' depends on a reliable phonetic transcription. Additionally, the definition of the phone as a unit is not unproblematic (see section 2.2.1); the same is true for the realised syllable (e.g. potentially syllabic consonants). The definition of the word is problematical too, e.g. because of compounding. Since the definition of some of the units affects the quantification of articulation rate, differently defined speech rate units make a comparison to other studies and other languages more difficult.

Apparently, realised syllables and phones correlate slightly higher with duration than their intended counterparts do. Although phone and syllable deletions are common in fast speech and result in a lower *measured* articulation rate, it is not clear what the effect on the *perceived* articulation rate is[2]. The perceived articulation rate may depend either on deletions and replacements of the intended units (see section 3.2) which complicate the lexical access and/or on temporal and spectral reductions which lead to a worse fit of the acoustic models (see section 4).

We conclude that the 'realised phone' best expresses the articulation rate. For this reason, we shall use this as the linguistic unit for measuring articulation rate in the rest of this article. In studies in which no realised phone labels are available, the 'intended syllable' should be given preference over graphical word. Both are easily derivable, but the graphical word showed an especially low correlation with duration. Although the 'intended phone' has an even higher correlation with duration than the 'intended syllable', its definition is more problematical.

With respect to the utterance domain the inter-pause stretch (ips) seems more appropriate than the intonation phrase (IP). The ips' correlation values score better in all comparisons, and they are easier to determine than intonation phrases which require some kind of prosodic annotation (not often available). Additionally, definitions of prosodic events and criteria to label them can differ considerably between studies.

3. Spontaneous versus Read Speech

We expect that spontaneous speech is marked by more changes in articulation rate than we find in read speech: Planning problems are likely to cause hesitations (e.g. syllable drawls) leading to slow stretches followed by fluent, fast stretches. These planning problems in spontaneous speech also increase the number of filled and unfilled pauses which lead to shorter ips. Especially utterances consisting of only

[2] As an example, a "full, unreduced" realisation of the sentence in section 2.2.1 with 2 sec duration would result in an articulation rate of 5 real. syll/sec, whereas the "reduced" version with the same duration would have 4 real. syll./sec.

one or two discourse particles such as "ja" contribute to a high number of short but very slow ips. The last points would support the reported tendency that "the longer the utterance the faster its rate" [4,12,13]. Emphasis, which occurs more often in spontaneous speech, represents another factor for slowing down.

3.1. Rate characteristics

A comparison of the rate characteristics of read versus spontaneous speech (table 2) shows that in spontaneous speech inter-pause stretches as well as intonation phrases are shorter on average and show a greater variance than in read speech.

Table 2: Mean duration (in sec) and mean articulation rate (real. phones/sec) of ips and IP for spontaneous and read speech with standard deviations.

		total	duration		artic. rate	
		n	mean	sd	mean	sd
spont.	ips	3757	1.81	1.29	13.24	3.29
	IP	5784	1.17	0.73	13.18	3.75
read	ips	4871	1.98	1.03	13.06	2.03
	IP	6474	1.49	0.67	13.01	2.23

With respect to articulation rate spontaneous speech is slightly faster and shows a greater variance. Although faster on average, spontaneous speech features a high number of slow utterances. One reason lies in the large number of very short ips (<1 sec) in this speaking mode. Indeed, one and two word utterances are slower than the mean.

Intonation phrases are generally shorter than ips, but there is basically no difference in articulation rate.

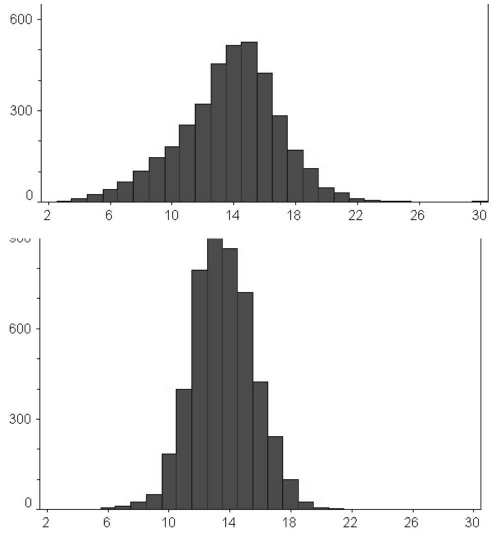

Fig. 2: Histograms of articulation rate (realised phones/sec per ips) in spontaneous (top) vs. read speech (bottom)

Correlating the duration of each ips with its rate the tendency towards higher articulation rates in longer utterances is not confirmed, neither for spontaneous (r=0.235) nor for read speech (r=0.116). The rather small r-values indicate that the articulation rate hardly increases in longer ips.

3.2. Deletions and replacements

Compared to fast speech, we expect fewer phone deletions and replacements in slower stretches. Reductions are classified either as deletions or as replacements of the lexical form. To evaluate this, all ips were grouped according to their articulation rate: "slow" < mean – 1sd; "medium" between mean +/– 1sd; "fast" > mean + 1sd.

The results in table 3 illustrate the increase of deletions and replacements from "slow" to "medium" to "fast". Although deletions are much more frequent than replacements this does not mean that all phones behave according to this pattern. For /n/ there are more replacements than deletions. In contrast, glottal stop /?/ and schwa /@/ show no replacements at all, but a very high number of deletions. The so-called "a-schwa" /6/ which is /@r/ underlyingly is seldom deleted or reduced.

Consonants are generally strongly affected by lexical alternations such as deletions and replacements, whereas vowels are not. Phone insertions are also found, but the number of insertions is negligible.

Table 3: Percentages of deletions and replacements of all phones as well as selected highly frequent phones for "slow", "medium" and "fast" read speech.

	deletions			replacements		
phone	slow	med	fast	slow	med	fast
all	10.1	13.1	16.2	0.9	1.4	2.4
n	1.4	2.4	5.0	7.2	9.8	9.7
?	52.1	61.7	76.8	0.0	0.0	0.0
@	36.0	43.3	51.6	0.0	0.0	0.0
6	1.0	0.5	1.3	2.4	0.7	0.4

3.3. Discussion

As expected spontaneous speech shows more slow and more short stretches than read speech. The greater variance in the articulation rate reflects the greater speech rate dynamics in spontaneous speech on a *global* level. These measures do not reflect more *local* variations, e.g. caused by an increased number of pauses and phrase boundaries (phrase-final lengthening), as well as more emphatic pitch accents (accentual lengthening) and dysfluent syllable drawls.

In line with our expectations we find an increasing number of segmental reductions with increasing articulation rate. Although this tendency is quite clear, the picture of the different phones and phone classes must be differentiated. We observed hardly any deletions and replacements of vowels (except schwa), so that if there is any vowel reduction, it must take place on the acoustic rather than on the lexical level.

4. Phone Classification

According to figure 1 we should expect a good classification for medium-rate phones, and lower classification rates for fast and slow phones. A possible explanation for this pattern is that the majority of the phones fall in the medium articulation rate category and therefore dominate the acoustic models.

4.1. Methods

The HMM ToolKit was used to train three-state left-to-right hidden Markov models (five states for diphthongs) with 8

Gaussian mixtures per state for each phone. Phone classification results were computed using a jackknife method (with five subexperiments) both for read and spontaneous speech.

4.2. Results

As in section 3 we shall only present results for ips. Averaged results for each jackknife experiment are shown in figure 3.

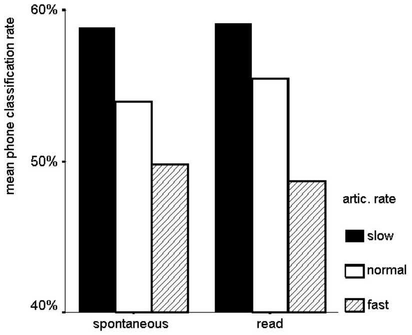

Fig. 3: Mean phone classification rates for "slow", "medium" and "fast" inter-pause stretches in spontaneous and read speech.

Articulation rate clearly affects the phone recognition rates: the differences between slow, normal and fast speech are highly significant in all cases (multiple *t*-tests for matched data). Although the task of classifying pre-segmented phones is not directly comparable to the word recognition experiments reported in [1-6], it is interesting to note that we find a drop in recognition rate with articulation rate, while many others have found lower recognition rates for fast *and slow* speaking rates [1,3,6]. The latter pattern is reflected in our experiments only for a few short vocalic phones.

In general, the recognition rates are higher for consonants than for vowels. Particularly, voiceless obstruents show high recognition rates, while diphthongs, /6/ and especially schwa are poorly recognised.

4.3. Discussion

For fast speech the expected lower phone classification score was indeed found. Contrary to our expectations, though, correct phone classification for slow speech was significantly higher than for medium-rate speech. A possible phonetic explanation for this finding is that slow phones are pronounced more clearly and are therefore easier to recognise. This tendency may be supported by the greater number of pitch-accented syllables in slower speech [14].

Although vowels, except schwa, were hardly affected by deletions and replacements (see section 3.2) the lower classification rate for vowels gives rise to the assumption that vowels are reduced on the acoustic level.

5. General Discussion

We first examined the optimal linguistic unit and domain to reflect temporal variance in speech. Although the 'realised phone' per inter-pause stretch showed the highest correlation with duration, other easily derivable units are good candidates to express articulation rate. The greater temporal dynamics in spontaneous compared to read speech is characterised by a higher number of short utterances and also a higher number of slower utterances despite the faster mean articulation rate. Surprisingly, these greater dynamics have no effect on the acoustic level in terms of phone classification differences.

One of the most interesting findings of the paper is the difference between phones in the type of changes that they undergo for different articulation rates. Consonants were found to be particularly sensitive to deletions and/or replacements on the lexical level, while vowels were more sensitive to reductions on the acoustic level, leading to low phone classification rates. The clearest exception to this rule is schwa, which shows strong reductions on both levels. Since schwa has a very high frequency in German, the results stress the importance of adding pronunciation variants to the lexicon of ASR systems. Also, the low phone classification rates for both /@/ and /6/ show that phonetic research is needed to improve the acoustic modelling. Particularly trans-consonantal co-articulation is likely to cause great variance which human listeners normalise for on the basis of the larger context which was missing in the classification task.

Acknowledgements

We would like to thank Caren Brinckmann for her help in generating the data analysed in section 3.

6. References

[1] Siegler, M. A. & Stern, R. M. "On the effects of speech rate in large vocabulary speech recognition systems." *Proc. ICASSP* Detroit (1), 612-615, 1995.

[2] Mirghafori, N., Fosler, E. & Morgan, N. "Fast speakers in large vocabulary continous speech recognition." *Proc. Eurospeech* Madrid. 1995.

[3] Brøndsted, T. & Printz Madsen, J. "Analysis of speaking rate variations in stress-timed languages." *Proc. Eurospeech* Rhodes, 481-484, 1997.

[4] Martínez, F., Tapias, D., Álvarez, J. & León, P. "Characteristics of slow, average and fast speech and their effects in large vocabulary continous speech recognition." *Proc. Eurospeech* Rhodes, 469-472, 1997.

[5] Pfau, T. & Ruske, G. "Creating Hidden Markov Models for fast speech." *Proc. ICSLP* Sydney, 205-208, 1998.

[6] Alleva, F., Huang, X., Hwang, M-Y. & Jiang, L. "Can continous speech recognisers handle isolated speech?" *Speech Communication* 26 (3), 183-190, 1998.

[7] Samudravijaya, K., Singh, S.K., Rao, P. "Pre-recognition measures of speaking rate." *Speech Communication* 24, 73-84, 1998.

[8] Kohler, K., Pätzold, M, & Simpson, A. "From scenario to segment. The controlled elicitation, transcription, segmentation and labelling of spontaneous speech." *Arbeitsberichte Phonetik Kiel* 29, 1995.

[9] http://www.phon.ucl.ac.uk/home/sampa/german.htm

[10] Miller, J.L., Grosjean, F. & Lomanto, C. "Articulation rate and its variability in spontaneous speech: a reanalysis and some implications." *Phonetica* 41, 215-225, 1984.

[11] Goldman-Eisler, F. "The significance of changes in the rate of articulation." *Lang. and Speech* 4, 171-174, 1961.

[12] Fónagy, I. & Magdics, K. "Speed of utterance in phrases of different lengths." *Lang. & Speech* 3, 179-192, 1960.

[13] Malécot, A., Johnston, R. & Kizziar, P.A. "Syllabic rate and utterance length in French." *Phonetica* 26, 235-251, 1972.

[14] Trouvain, J. & Grice, M. "The effect of tempo on prosodic structure." *Proc. 14th ICPhS*, 1067-1070, 1999.