INTERSPEECH 2009 BRIGHTON

Word-final [t]-deletion: An analysis on the segmental and sub-segmental level

Barbara Schuppler[1], *Wim van Dommelen*[2], *Jacques Koreman*[2], *Mirjam Ernestus*[1,3]

[1]Center for Language and Speech Technology, Radboud University Nijmegen, The Netherlands
[2]Department of Language and Communication Studies, NTNU, Trondheim, Norway
[3]Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands

b.schuppler@let.ru.nl, {wim.van.dommelen, jacques.koreman}@hf.ntnu.no, mirjam.ernestus@mpi.nl

Abstract

This paper presents a study on the reduction of word-final [t]s in conversational standard Dutch. Based on a large amount of tokens annotated on the segmental level, we show that the bigram frequency and the segmental context are the main predictors for the absence of [t]s. In a second study, we present an analysis of the detailed acoustic properties of word-final [t]s and we show that bigram frequency and context also play a role on the sub-segmental level. This paper extends research on the realization of /t/ in spontaneous speech and shows the importance of incorporating sub-segmental properties in models of speech.
Index Terms: Acoustic Reduction, Phonetic Detail, Automatic Transcription, Pronunciation Variation

1. Introduction

A frequent phenomenon observed in spontaneous, conversational speech is that words are produced in a reduced way compared to a transcription of their canonical pronunciation; a phrase like 'was supposed to see' may sound just like [səsəsi]. A study on American English shows that deletions of whole syllables occur in 6% of the word tokens and that segment deletions and alternations occur in even every fourth word [1]. Investigations of the conditions under which these phenomena are likely to occur are of importance for building psycholinguistic models of speech perception and production and for improving automatic speech recognition (ASR) systems.

Several studies have investigated the reduction of [t] in Germanic languages. For instance, Jurafsky et al. [2] investigated predictors for the acoustic presence of [t]s in conversational American English. For other languages however, quantitative studies on [t]-deletion have focused on dialectal variations [3], but not on standard speech. Moreover, investigations so far have only considered the segmental level for investigating the reduction of /t/. Mitterer and Ernestus [4] reported variation on the sub-segmental level, but they only described what kind of sub-segmental variation can occur and which cases they label as absent and which as present. The aim of this paper is to investigate the conditions which favour reduction of [t]s in conversational standard Dutch on both the segmental and the sub-segmental level.

Well-studied predictors for reduction in general are the lexical frequency of the word [5], with more reduction in high-frequency words, and its lexical class, with function words being more reduced than context words [6]. Bell et al. [6] report that for function words the bigram frequency with the previous word is a predictor for more reduced articulation, whereas for content words the bigram frequency with the following word is of more importance. Furthermore, it has been shown that segmental context plays an important role. More /t/-deletions are observed when /t/s are preceded and followed by consonants than by vowels [7, 4]. Finally, also morphology can predict reduction. For instance, Pluymaekers et al. [5] showed that the presence of a morphological boundary affects the duration of segments.

This paper presents two studies. First, we investigated the presence versus absence of [t]s on the basis of a large corpus of conversational Dutch, for which a phonemic transcription has been generated using an ASR system. Second, we will present an investigation of several sub-segmental acoustic properties of word-final [t]s, based on manual annotations of a small part of the material used in the first study.

2. Corpus Data

Our research is based on six spontaneous Dutch dialogues that are part of the CORPUS ERNESTUS [7]. Each of the conversations has a length of approximately 90 minutes. In total, the six conversations contain 89716 word tokens and 6296 word types produced in 8.5 hours of speech. Characteristic for this corpus is the high level of spontaneity and the homogeneity in geographical and social background of the speakers. All 12 speakers were male native speakers of Dutch, they all lived in the western provinces of the Netherlands and had academic degrees. The speakers were between 21 and 55 years old. They have been classified as speakers of standard Dutch by trained phoneticians [7].

3. Study 1

3.1. Material

For the six dialogues, we generated a phonemic transcription by using an ASR system. Automatic transcriptions have the advantage that they are consistent and can be easily obtained for large data sets. The hand-made orthographic transcription of the six dialogues was prepared for automatic processing [8], and a forced alignment was carried out by means of the speech recognition toolkit HTK [9]. For this alignment, 37 32-Gaussian tri-state monophone acoustic models [10] were trained on the Dutch library of the blind of the CGN (Corpus Gesproken Nederlands) [11]. The lexicon used for the alignment contained canonical phonemic representations and several pronunciation variants for each word type. These variants were generated by applying a set of 32 reduction rules to the canonical forms of the words. These rules were formulated on the basis of observed reduction processes from earlier studies on spontaneous, casual Dutch [7] and included one rule that deleted word-final [t]s, independent of segmental context. The rules created on average

Factor	β	z-value	p-value
M1: On all data	$N =$	8270	
Intercept M1	0.55	2.76	$<$.0.001
Bigram-Frequency	-0.62	-5.78	$<$.0001
Previous-Segment: vowel	1.23	7.86	$<$.0001
Following-Segment: vowel	1.164	9.52	$<$.0001
Following-Segment: silence	1.02	11.02	$<$.0001
Word-Class	-0.35	-1.05	$<$1
Bigram-Frequency:Word-Class	0.32	2.20	$<$.01
M2: [t] preceded by consonant	$N =$	2689	
Intercept = fricative	-0.73	-2.44	$<$.0.01
Bigram-Frequency	-0.38	-2.69	$<$.001
Previous-Segment: glide	2.79	4.59	$<$.0001
Previous-Segment: liquid	1.71	7.83	$<$.0001
Pirvious-Segment: nasal	1.26	5.44	$<$.0001
Previous-Segment: plosive	0.81	3.11	$<$.001
M3: [t] followed by consonant	$N =$	4922	
Intercept = fricative	0.89	3.40	$<$.0001
Bigram-Frequency	-0.63	0.13	$<$.0001
Following-Segment: liquid	1.19	4.13	$<$.0001
Following-Segment: plosive	-0.95	-7.03	$<$.0001
Following-Place: homorganic	-0.61	-5.16	$<$.0001
Bigram-Frequency:Word-Class	0.38	2.13	$<$.01

Table 1: *Study 1: Statistical Summary. Intercept M1: preceding consonant and following consonant for content content words*

27.06 pronunciations per word type.

We based our analysis of word-final /t/s on 8270 word tokens representing 556 word types. These tokens were chosen in such a way that the following word was part of the same syntactic phrase. We excluded the highly frequent words *het* 'it' and *heeft* 'has', since they were represented by a much higher number of tokens (i.e., 1511 and 96, respectively) than the other words (average number of tokens: 14.83) and because of these extremely high frequencies may show idiosyncratic behavior.

3.2. Results and Discussion

We observed that 35.15% of all word-final /t/s were classified as absent. To investigate the conditions favoring the presence of word-final [t]s we used the statistical modeling technique of mixed-effects logistic regression with a logit link function and contrast coding [12]. This model predicts $log(\frac{p}{1-p})$ with p the chance that the [t] is present. In all models that we will present in this section, i.e., M1, M2, M3, we included the random variables Speaker, Word, and Following-Word, because they appeared statistically significant predictors ($p < .0001$ for these three random variables). The independent variables were Previous-Segment and Following-Segment, where the former can have the values Consonant and Vowel and the latter can have the additional value Silence. Furthermore, the models included the independent variable Word-Class, with the two values Function Word and Content Word, Number-of-Syllables (values between 1 and 6), the Syllabic-Stress, which indicates whether the word-final syllable is stressed, and the Morphological-Class. This class distinguished cases where the /t/ is a grammatical morpheme by itself (i.e., indicating the second or third person singular present tense as in *loop-t*, 'run-s') from those were it is part of the stem (e.g. as in *kast*, 'cupboard'). Finally, we included the independent variables Word-Frequency and Bigram-Frequency, where we define Bigram-Frequency as the frequency of the target word with the following word. Both frequency measures were extracted from the CGN [11]. The logged Word-Frequency has a maximum of 12.60 and the logged values for the Bigram-Frequency are between 0.52 and 2.41. Since the two measures are highly correlated ($r = .61, p < .0001$) and the Bigram-Frequency showed a distribution closer to a normal distribution, we first included only the Bigram-Frequency in the models. Predictors and interactions that did not show statistically significant effects were removed from the models. Below we only report the significant effects.

Table 1 shows the results for the first model (M1), which was based on the complete data set. Both Previous and Following-Segment showed to be significant: [t]s are less often absent after vowels (30.0%) than after consonants (48.05%) and they are less often absent before vowels (25.9%) and before silence (24.20%) than before consonants (43.05%). The effect of Bigram-Frequency shows that word-final [t]s were more often absent in word combinations of a higher frequency. Separate analysis of the function words and content words revealed that this effect was significant for both content and function words but much greater for content words ($\beta = -0.62, z = -5.37, p < .0001$ for content words versus $\beta = -0.31, z = -2.66, p < .001$ for function words). In order to compare the contributions of Bigram-Frequency and Word-Frequency, we refitted model M1, replacing Bigram frequency by Word-Frequency. Word-Frequency did not show an interaction with Word-Class. Its main effect was statistically significant, but its effect size ($\beta = -0.11$) was smaller than the effect size of Bigram-Frequency ($\beta = -0.62$). In a next step, we orthogonalized Word- Frequency and Bigram-Frequency by replacing Word-Frequency by the residuals of the linear regression model predicting Word-frequency as a function of Bigram frequency. We added these residuals in model M1 also in interaction with Word-Class. The results indicate that these residuals do not have any predictive power ($p < .1$). We therefore conclude that it is especially bigram frequency that predicts the presence of word-final [t].

For the subgroup tokens where the /t/ is preceded by a consonant, we built another model (M2) in order to test the effects of the place and manner of articulation of this consonant. Place-of-Articulation could be either homorganic or heterorganic with the place of articulation of the [t], which is alveolar for Dutch. Manner-of Articulation had the values fricative, plosive, nasal, glide and liquid. These variables were added to the predictors mentioned above after exclusion of Previous-Segment. Table 1 shows the values of the significant variables of this model (M2). [t]s are absent most when preceded by a fricative (60.0%), with significantly more deletions than after plosives (51.1%) and highly significantly more deletions than after nasals (48.9%), liquids (32.3%) and after glides (20.6%). To find out whether there were also significant differences between plosives, nasals, liquids and glides we ran the same model on the data excluding the fricatives. The difference between glides and liquids is not significant, but there are significantly fewer [t]s after nasals ($\beta = -1.43, z = -2.58, p < .001$) and plosives ($\beta = -1.71, z = -3.014, p < .001$) than glides.

For the subgroup of tokens where the /t/ is followed by a consonant, we built a third model (M3) to investigate the effects of the place and manner of articulation of this consonant. In addition to the independent variables mentioned above, we included Place-of-Articulation and Manner-of-Articulation with the exclusion of Following Segment. We observe significantly fewer present [t]s when followed by a homorganic consonant. Furthermore, [t]s are absent significantly less often be-

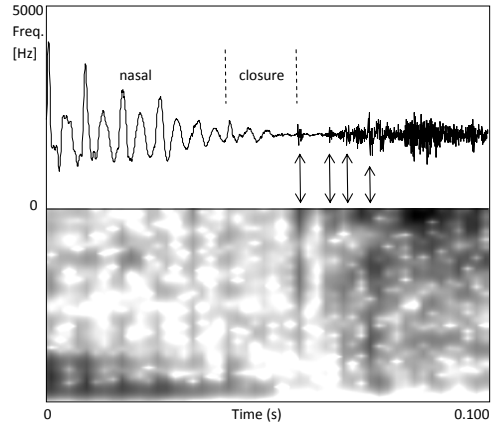

Figure 1: *Waveform and spectrogram of /ntz/ in [hɑnt sʌu]. The closure is filled with nasal friction, the multiple weak bursts are indicated by arrows.*

fore liquids (19.7%) but more often before plosives (57.8%), while there was no difference between fricatives (39.9%), nasals (39.4%) and glides (31.2%). Plosives that are homorganic with /t/ are /t/ and /d/, which could mean that the results are a mere proof of the well known degemination. We excluded this segmental context and ran the model again. Since we found that both variables were still significant (Place-of-Articulation: Homorganic: $\beta = -0.25, z = -2.07, p < .01$; Manner-of-Articulation: plosive: $\beta = -0.82, z = -5.35, p < .0001$) we conclude that [t]s are less often present before homorganic consonants in general and before all kinds of plosives. Possibly /t/s are more often absent before plosives due to gestural overlap [13].

4. Study 2

4.1. Material and Method

For this study we used a set of 130 word tokens representing 65 word types which were a subset of the tokens analyzed in Study 1. The chosen tokens were either preceded by a vowel or a homorganic nasal and followed by a fricative in the following word. One third consisted of function words, one third of content words where the word-final /t/ was part of the stem and the last third were verb forms in which the /t/ was the suffix indicating the second or the third person singular of the present tense, e.g. *loop-t* ('walk-s').

First, the word-final /t/s of the target words were classified as perceptually present or absent by consensus of at least two of the authors, one a native speaker of Dutch, the other a native speaker of German. For the sub-segmental analysis of the word-final /t/s, we indicated whether a closure was completely absent or there was some acoustic exponent present in the signal, for example low-amplitude friction, nasality, voicing or silence. With respect to the burst we specified whether there was none, one, or more than one. In addition, we labeled them as strong or weak, where weak bursts were characterized as extremely short and with energy only in part of the spectrum. Since the following segment was always a fricative, all target /t/s were followed by friction. We specified whether alveolar friction was present and whether it was voiced or not. Furthermore, we annotated whether the friction started with an abruptly rising amplitude or smoothly and if abruptly, whether it started simultaneously with the burst. Figure 1 shows an example for a /t/ in [hɑnt sʌu] ('hand shall'). The realization of the word-final /t/ was annotated as having a closure filled with nasal friction, multiple weak bursts and a simultaneous smooth start of voiceless alveolar friction with the second burst.

Property	Present	Absent	Details		
Closure	Present	Absent	Silence	Frict.	Nasal
	86 (25)	44	57 (13)	20 (6)	9 (6)
Burst	Present	Absent	Multiple		
	75 (39)	55	22 (9)		
Abr.frict.	Present	Absent	W. burst		
	83	47	63		
Alv.frict.	Present	Absent			
	60 (7)	14			

Table 2: *Counts of acoustic obervations. Abr.frict. = abruptly starting friction; W. burst = Abr.frict. starting with the burst. Alv.frict. = alveolar friction; In brackets: voiced cases for present closures and Alv.frict. or strong cases for present bursts.*

4.2. Results and Discussion

4.2.1. Variation on the sub-segmental level

Table 2 presents the counts of all the acoustic observations annotated for the set of target words. This table shows how much variation there is on the sub-segmental level, even for this material of limited segmental context. 'Canonical' realizations of /t/s, i.e. containing a voiceless closure, one strong burst, and produced with an alveolar articulation place, occur in only 11.5% of our data and complete absence of any /t/ properties is even less frequent (5.4%).

For investigating which variables predict sub-segmental acoustic properties, we built statistical models using mixed effects logistic regression, as for Study 1, with Word and Speaker as the crossed random variables. In each model, the independent variables were the sub-segmental properties (Closure, Burst, Alveolar-Friction, Abrupt-Friction), except if that property was the dependent variable. Further independent variables were Previous-Segment, Following-Segment, Word-Class and Morphological-Class. Finally, we included the logged Word-Frequency (min = 0.69, max = 12.08) and Bigram-Frequency (min = 0, max = 2.09). These two measures are highly correlated ($r = 0.59, p < .0001$), so we used Word-Frequency only because it showed a distribution closer to a normal distribution. Predictors and interactions that did not show statistically significant effects were removed from the models and we thus only report the significant effects.

The first statistical model investigated the type of closure. The dependent variable Type-of-Closure had the two values Silence and Filled, which could be either low amplitude friction, nasality, or voicing. We observed that significantly more frequent words tend to be produced less often with completely silent closures ($\beta = .20, z = 2.25, p < .05$). Filled closures typically result from coarticulation with preceding segments and the result is therefore in line with the finding that more frequent words tend to be more reduced.

Then, we investigated the properties of the burst in a model with Burst (yes or no) as the dependent variable. Word-Frequency showed to be a significant ($\beta = -0.81, z = -2.07, p < .01$) predictor: Bursts are less probable to be

present in words of a higher frequency. Again this shows that higher frequent words tend to be more reduced. Also the presence of a Closure was a significant predictor: ($\beta = 2.15, z = 4.53, p < .0001$). Bursts are more likely if also a closure is present. Next, we investigated the presence of one versus more than one bursts. Multiple bursts appeared more likely ($\beta = 2.25, z = 3.65, p < .0001$) when the word-final /t/ is followed by a homorganic than by a heterorganic fricative.

In order to investigate the predictors for the existence of Alveolar-Friction, we looked at only those tokens where the following segment was [s] nor [z], since in those cases the [t] is followed by alveolar friction anyway. The independent variable Previous-Segment had a significant effect: Alveolar friction occurs more often when the /t/ of interest is preceded by a vowel ($\beta = 1.73, z = 2.57, p < .01$).

4.2.2. ASR Classification vs. Auditory Classification

On the basis of the material of Study 2, we compared the classification of the word-final /t/s as perceptually present or not with the automatically generated transcription. Overall, there was an agreement on 75.4% of the cases, which is quite high compared with data on agreement between human transcribers, which reached 78.8% for spontaneous speech [14]. These high deviations between human transcribers are caused by the difficulty of transcribing spontaneous speech, where transcribers tend to hear intended segments.

Whereas perceptually only 25.4% of all /t/s were classified as absent, 40.8% of [t]s were absent in the automatic transcriptions. Given the low proportion of canonical [t]s that we found, this should perhaps not come as a surprise, the more so because the HMMs were trained on read speech [11]. Our sub-segmental analysis of [t] suggests that automatic transcriptions tools would be improved if they first identify the intervals where a [t] might be realized (given the canonical representation of a word) and then apply detailed analysis and classification techniques informed by data such as those in Table 2.

5. General Discussion and Conclusions

This paper presents an analysis of the realization of word-final /t/s based on a large corpus of conversational standard Dutch, considering both the segmental and sub-segmental level.

In our study on the absence versus presence of [t]s, we could show similar effects of bigram frequency for our Dutch material as Bell et al. [6] did for English: [t]s are absent more often when the bigram frequency of the target word with the following word is high and this effect is stronger for content words. Secondly, we observed that segmental context plays an important role for the realization of /t/s. Our results are in line with previous reports that [t]s are mainly absent in consonant clusters [7, 4]. What is more, we present a consistent quantitative analysis for all segmental contexts.

In a second study, we extended previous research on the realization of /t/ by analyzing the sub-segmental level. For the large space between 'canonically present' (11.5% of cases) and 'completely absent '(5.4%), we showed what kind of variation occurs and how often. When analyzing the conditions for the sub-segmental acoustic cues, we found that a high lexical frequency predicts the absence of bursts and that closures are more often filled with friction or nasality. These findings are in line with the generalization that more frequent words tend to be more reduced. Also on this level the context has a big impact. We saw that alveolar friction occurs more often when the [t] follows a vowel and that more multiple bursts occur before homorganic consonants.

In conclusion, the present study supports previous findings on the realization of /t/ in casual speech on the basis of a large corpus of spontaneous Dutch [6, 2]. In addition, it has documented variation on the sub-segmental level and has shown that also at this level more frequent words tend to be more reduced. Psycholinguistic models of speech production and comprehension have to take this variation into account in order to explain the processing of every day speech. Similarly, it may be possible to improve ASR systems by including sub-segmental information into account of the type documented in this paper.

6. Acknowledgements

This research was supported by the Marie Curie Project "Sound to Sense". Mirjam Ernestus was supported by a EURYI-award from the European Science Foundation.

7. References

[1] K. Johnson, "Massive reduction in conversational American English," in *Spontaneous Speech: Data and Analysis*, K. Yoneyama and K.Maekawa, Eds. Tokyo, Japan: The National International Institute for Japanese Language, 2004, pp. 29–54.

[2] D. Jurafsky, A. Bell, M. Gregory, and W. D. Raymond, "Probabilistic relations between words: Evidence from reduction in lexical production," in *Frequency and the emergence of linguistic structure.*, J. Bybee and P. Hopper, Eds. John Benjamins, 2001, pp. 229–254.

[3] T. Goeman, "T-deletie in nederlandse dialecten. kwantitatieve analyse van structurele, ruimtelijke en temporele variatie." Ph.D. dissertation, HAG, 1999.

[4] H. Mitterer and M. Ernestus, "Listeners recover /t/s that speakers reduce: Evidence from /t/-lenition in Dutch," *Journal of Phonetics*, vol. 34, pp. 73–103, 2006.

[5] M. Pluymaekers, M. Ernestus, and H. R. Baayen, "Lexical frequency and acoustic reduction in spoken Dutch," *The Journal of the Acoustical Society of America*, vol. 118, no. 4, pp. 2561–2569, 2005.

[6] A. Bell, J. M. Brenier, M. Gregory, C. Girand, and D. Jurafsky, "Predictability effects on durations of content and function words in conversational English," *Journal of Memory and Language*, vol. 60, pp. 92–111, 2009.

[7] M. Ernestus, "Voice assimilation and segment reduction in casual Dutch. A corpus-based study of the phonology-phonetics interface," Ph.D. dissertation, LOT, 2000.

[8] B. Schuppler, M. Ernestus, L. Boves, and O. Scharenborg, "Preparing a corpus of Dutch spontanteous dialogues for automatic phonetic analysis." Interspeech, 2008, pp. 1638–1641.

[9] S. Young, G. Evermann, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, "The HTK book (version 3.2)," Cambridge University. Engineering Department., Tech. Rep., 2002.

[10] A. Hämäläinen, L. ten Bosch, and L. Boves, "Modelling pronunciation variation using multi-path hmms for syllables." ICASSP, 2007.

[11] N. Oostdijk, W. Goedetier, F. V. Eynde, L. Boves, J.-P. Martens, M. Moortgat, and H. Baayen, "Experiences from the spoken Dutch corpus project." LREC, 2002, pp. 340–347.

[12] T. F. Jaeger, "Categorical data analysis: Away from anovas (transformation or not) and towards logit mixed models," *Journal of Memory and Language*, vol. 59, pp. 434–446, 2008.

[13] C. Brownman and L. Goldstein, "Articulatory phonology: An overview." *Phonetica*, vol. 49, pp. 155–180, 1992.

[14] A. Kipp, M. Wesenick, and F. Schiel, "Pronunciation modeling applied to automatic segmentation of spontaneous speech." Eurospeech, 1997, pp. 1023–1026.