Predicting human perception and ASR classification of word-final [t] by its acoustic sub-segmental properties

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

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

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

Abstract

This paper presents a study on the acoustic sub-segmental properties of word-final /t/ in conversational standard Dutch and how these properties contribute to whether humans and an ASR system classify the /t/ as acoustically present or absent. In general, humans and the ASR system use the same cues (presence of a constriction, a burst, and alveolar friction), but the ASR system is also less sensitive to fine cues (weak bursts, smoothly starting friction) than human listeners and misled by the presence of glottal vibration. These data inform the further development of models of human and automatic speech processing.

Index Terms: Sub-segmental Acoustic Properties, Automatic Transcription, Human speech Perception, Dutch

1. Introduction

In conversational speech words are often realized in a reduced way compared to their citation forms. For instance, in conversational Dutch, 35 % of the word-final [t]s are absent: A word like *niet* 'not' may be realized as [ni] [1]. Reduction may also imply that segments are produced in overlap with surrounding segments, resulting in variation at the sub-segmental level. In an earlier study, we documented what kind of variation occurs at the sub-segmental level for word-final /t/. Importantly, we have shown that in a corpus of conversational Dutch only 11.5% of /t/s were realized canonically, while complete absence of all sub-segmental properties for /t/ was observed in only 5.4% of the cases [1]. Thus, 83.1% of the /t/s had only sub-segmental properties present in the signal. Listeners are used to deal with gradient reduction in every day conversations. For instance, Janse et al. [2] showed that listeners are capable of distinguishing fully released, unreleased and deleted [t]s. The first aim of the present paper is to investigate on the basis of which sub-segmental properties listeners classify word-final /t/ as acoustically present.

These data will inform psycholinguistic models of speech comprehension that aim to account for pronunciation variation, including variation from reduction. Scharenborg [3] showed that the performance of existing psycholinguistic models of speech perception (e.g. Shortlist [4]) is improved by the incorporation of information about pronunciation variation at the segmental level in the models' lexicons. She also showed that performance is improved by the incorporation of durational information [5]. However, to our knowledge, no model has yet explicitly taken into account variation at the sub-segmental level[1].

[1]It might be argued that episodic models do take sub-segmental properties into account, but then again without making these explicit.

Similarly, Automatic Speech Recognition (ASR) systems have been improved for spontaneous speech by the incorporation of pronunciation variation at the segmental level [6, 7]. There is work on ASR systems that are based on phonetic features in order to capture coarticulation effects (e.g. [8], [9], [10]), but these systems are progressing slowly. One might also think of ASR systems that make use of sub-segmental information in addition to segmental information, similar to what humans seem to do. The second aim of the present paper is to investigate on the basis of which sub-segmental properties current ASR systems that operate on the segmental level (fail to) classify word-final /t/ as acoustically present or absent. Comparisons of these cues with those used by human listeners may show us how to improve ASR systems.

The rest of this paper is organized as follows. Section 2 describes the corpus material and the three types of annotations that we made (automatic phonetic transcriptions, perceptual classification and manual annotation of sub-segmental properties). Section 3 analyzes which sub-segmental properties predict the perceptual presence of word-final /t/ for human listeners. In Section 4, we first analyze which sub-segmental properties condition the ASR classification of the word-final /t/s as present versus absent and secondly compare the results with those from human perception. On the basis of this comparison, we provide suggestions of how to improve automatic transcription systems. The paper ends with a summary of the findings.

2. Material and Method

2.1. Corpus Data

The present study is based on the ten spontaneous dialogues of the ERNESTUS CORPUS OF SPONTANEOUS DUTCH [11], which together contain 153,200 word tokens and 9,035 word types produced in 15 hours of speech. Characteristic for this corpus is that the 20 speakers form a homogeneous group with respect to their geographical and social background. Since the speakers were friends speaking about everyday issues, the atmosphere during the conversations was relaxed, resulting in a casual, chatty speech style.

From this corpus, we extracted 486 word tokens (mono and multi-syllabic) representing 141 word types ending in [t] in their citation form. Since final devoicing is a characteristic of Dutch, this set of words contains both words that orthographically end in *-d* and *-t*. The tokens were taken from a limited number of segmental contexts: The /t/ was preceded by a vowel or by /n/ (which has the same place of articulation as the /t/) and followed by a word starting with either a vowel, a fricative, or a plosive.

Property	Present	Absent	Details	
Constriction	Present	Absent		
	392 (77)	94		
Burst	Present	Absent	Multiple	One
	254 (115)	232	102 (51)	152(63)
Start fri.	Smooth	Abrupt	Simultan.	
	239	114	108	
Alv.fri.	Present	Absent		
	391 (28)	95		

Table 1: *Counts of acoustic observations. Start fri. = smoothly versus abruptly starting friction of the following segment; Simultan. = Friction is starting simultaneously with the burst. Alv.fri. = alveolar friction; In parentheses: voiced cases for "Constriction Present" and "Alv.fri. Present"; weak cases for "Burst Present", as opposed to strong bursts.*

Further, it contained equal numbers of tokens from the speakers and both function and content words.

2.2. Annotations

2.2.1. Automatic Phonetic Transcriptions

We created automatically a phonemic transcription for the ERNESTUS CORPUS OF SPONTANEOUS DUTCH by means of a forced alignment using the toolkit HTK [12]. Input for the forced alignment are the speech files, the orthographic transcriptions, a pronunciation lexicon linking the orthographic transcriptions with phonemic representations, and acoustic phone models.

The lexicon contains for each word its canonical representation and several pronunciation variants, which were generated by means of 32 reduction rules applied to the canonical pronunciation [7]. These rules include one that deletes [t] in word final position, independently of any other criteria and one that creates a variants where the word-final /t/ was realized voiced. The lexicon contains on average 27.06 pronunciations per word type.

The ASR system looks up each word from the orthographic transcription in the lexicon, and chooses the pronunciation variant that matches best with the speech signal, given the acoustic models. The acoustic models were 37 32-Gaussian tri-state monophone acoustic models [13] that had been trained on the Dutch library for the blind [14]. The models were trained at a frame shift of 5 ms, instead of the default of 10 ms, in order to transcribe short segments more accurately.

2.2.2. Human Perception: Auditory Classification

Two phoneticians, both native speakers of Dutch listened closely to all 486 selected words tokens in their context taking into account variably long stretches of speech. They classified their final /t/ by consensus as either acoustically strong, weak, assimilated, or absent. This classification appeared to be a difficult task, because stretches of truly conversational speech are often not literally understandable. Moreover, the recordings contained background noises and the distance between the speaker and the microphone sometimes varied strongly. Finally, some word tokens were produced at very high speech rates.

The two phoneticians classified 57.8% of the tokens as perceptually 'strong', 13.4% as perceptually 'weak', 7.0% as 'assimilated', and 21.8% as 'absent' (c.f. Table 3).

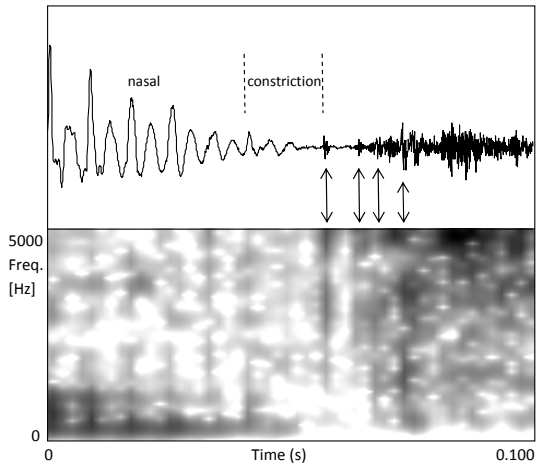

Figure 1: *Waveform and spectrogram of /ntz/ in [hant sʌu]. The multiple weak bursts are indicated by arrows.*

2.2.3. Annotation of Sub-segmental Properties

For each token, the same two phoneticians annotated by consensus the following sub-segmental properties of the [t]. First, they characterized a constriction as voiced, unvoiced, or absent. On the basis of the spectrograms, they annotated bursts as absent, as consisting of a single burst, or of multiple bursts. Present bursts were additionally specified as either strong or weak, with weak bursts having extremely short durations and energy only in part of the spectrum. For /t/s followed by an alveolar consonant, the friction was annotated as smooth if it showed gradually increasing amplitude, otherwise it was annotated as abrupt. We indicated whether this friction started simultaneously with the /t/ burst or not. Finally, for those tokens that were not followed by an alveolar obstruent (i.e. tokens that were not followed by /s/, /z/, /d/ or /t/), the phoneticians judged by ear whether alveolar friction was present. Figure 1 shows an example for a /t/ in [hɑnt sʌu] ('hand would'). The realization of the word-final /t/ was annotated as having a constriction, multiple weak bursts and a simultaneous smooth start of voiceless alveolar friction with the burst.

Part of the tokens (130 tokens) were annotated for an earlier study [1]. Table 1 shows how often each of the sub-segmental properties occur in our material.

3. Human Classification

The following analysis was motivated by two questions: Which are the sub-segmental properties that predict the perceptual presence of /t/ for humans? Which are the sub-segmental properties that distinguish between weak perceptual presence and complete absence?

3.1. Properties Predicting Presence

All statistical models in this paper are mixed-effects logistic regressions with a binomial logit link function and contrast coding [15]. For answering the first question, we built models predicting the perceptual presence of word-final [t], for which we merged /t/s classified as 'strong', 'assimilated' and 'weak'. Speaker ($p < .01$) was the only significant random variable. The speakers varied in the percentage of perceptually absent

/t/s between 6.7% and 42.9%. The independent variables were Constriction and Voicing (of the constriction), both with the values 'present' and 'absent', and Burst with the values 'absent', 'one' and 'multiple'. These sub-segmental properties have values for the complete data set. The other sub-segmental properties shown in Table 1 are included in models on subsets of the data. Predictors and interactions that did not show statistically significant effects were removed from the models.

As expected, both the presence of a constriction ($\beta = 2.97, z = 3.37, p < .0001$) and the presence of multiple bursts ($\beta = 1.57, z = 3.66, p < .0001$) or one single burst ($\beta = 1.62, z = 3.37, p < .0001$) predict the perceptual presence of /t/s. There was no statistically significant difference between stops with single and multiple bursts. A constriction was present in 93.1% of the /t/s classified as present and in 35.8% of the /t/s classified as absent. A burst was present in 62.6 % of the /t/s classified as present and in 15.9% of the /t/s classified as absent.

We carried out further analysis on those 384 tokens that were not followed by a homorganic fricative in order to investigate the role of Alveolar-Friction. All significant predictors of this model (*M1*) are shown in Table 2. In addition to the significant factors from the previous model, this model shows that Alveolar-Friction is a significant predictor for the perceptual presence of /t/s. Alveolar friction was present in 85.4% of the /t/s classified as present and in 34.2% of the /t/s classified as absent.

Next we built a model for those tokens that were followed by an obstruent and we saw that whether the following friction started smoothly or abruptly did not significantly predict the perceptual presence of /t/.

3.2. Properties of Weak Realizations

In order to answer the second research question concerning which variables predict acoustically 'weak' versus 'absent' /t/s, we built a model for the tokens that were classified as such ($N = 171$). We included the same independent variables and random factor in the model as we did for the first model. The resulting model only shows a significant main effect for the presence of a constriction ($\beta = 3.23, z = 4.91, p < .0001$). A constriction was present in 83.7% of the /t/s that were classified as perceptually weak and in 35.8% of the /t/s that were classified as absent.

We then focussed on those tokens that are followed by an obstruent, for which we have values on whether the friction started smoothly or abruptly ($N = 149$). We observed again that a perceptual weak presence is predicted by the presence of a constriction ($\beta = 2.10, z = 4.99, p < .0001$). In addition, [t]s are more likely to be classified as 'weakly' present if the friction starts abruptly ($\beta = 0.99, z = 2.44, p < .01$) rather than smoothly. This model (*M2*) is shown in Table 2.

4. ASR Classification

We analyzed which sub-segmental properties, as scored by the two phoneticians, predict the ASR classification of the word-final /t/s as present versus absent in the acoustic signal. Table 2 shows all significant predictors of this model (*M3*). The presence of a constriction and of multiple bursts or one single burst favor a /t/ to be labeled as present. A constriction was present in 92.2% of the /t/s classified as present and in 55.8% of the /t/s classified as absent. Similarly, a burst was present in 63.8% of the /t/s classified as present and in 29.2% of the /t/s classified as absent. Furthermore, Voicing appeared to be a significant predictor for the automatic classification: A voiced [t] was more often classified as absent (40.3%) than a voiceless [t] (17.5%).

Factor	β	z-value	p-value
M1: Perception: present			$N = 384$
Intercept: Burst: none	-1.80	-5.18	<.0001
Burst: one	1.12	2.14	<.01
Burst: multiple	1.18	1.97	<.01
Constriction: yes	2.64	7.02	<.0001
Alveolar Friction: yes	1.45	4.01	<.0001
M2: Perception: weak			$N = 149$
Intercept	-2.39	-5.19	<.0001
Constriction: yes	2.10	4.99	<.0001
Alveolar Friction: yes	0.99	2.44	<.01
M3: ASR			$N = 486$
Intercept	-1.11	-4.12	<.0001
Burst: one	0.99	3.34	<.0001
Burst: multiple	0.99	2.97	<.001
Constriction: yes	2.16	7.11	<.0001
Voicing: true	-0.89	-3.14	<.001

Table 2: *Statistical summary.*

Next, we built a model for those tokens that were not followed by a fricative homorganic with the /t/. The variables from the previous model were again significant. Moreover, we observed that significantly more /t/s were labeled as present if Alveolar-Friction was present ($\beta = 1.33, z = 4.05, p < .0001$). Alveolar frication was present in 81.3% of the /t/s classified as present and in 34.7% of the /t/s classified as absent.

Separate analysis of those tokens that were followed by an obstruent showed that whether the following friction starts smoothly or abruptly did not predict the presence of /t/. Finally, we investigated whether strong and weak bursts differ in how much they cue the presence of /t/ in all those tokens that were realized with a burst. We observed that less often /t/ was transcribed as present ($\beta = -1.02, z = -2.70, p < .001$) if the burst was weak (75.6%) than if it was strong (87.8%).

4.1. ASR Classification versus Human Classification

We compared the human classification of the word-final /t/s with the automatically generated transcription (Table 3). The agreement for the tokens that were classified as perceptually strong was 86.1% and for the tokens that were classified as perceptually absent 79.3%. The agreement for the perceptually more difficult tokens was lower: 66.2% of the perceptually

	%Total	**Forced Alignment** 'present'	'absent'	Agreement
%Total	-	68.3%	31.7%	-
Perception				
'strong'	57.8%	242	39	86.1%
'weak'	13.4%	43	22	66.2%
'assim.'	7.0%	25	9	73.5%
'absent'	21.8%	22	84	79.3 %

Table 3: *Perceptual classification vs. automatic annotation with forced alignment. 'assim.' = assimilation; %Total: Percent of all tokens that were classified as the given category.*

weak and 73.5% of the assimilated word-final /t/s were labeled as present by the ASR system. These numbers suggest that the overall agreement between the human and ASR classification is similar to the agreement among human transcribers, for example 78.8% for segmental transcriptions of spontaneous speech reported by [16, 17]. High deviations between human transcribers result from the difficulty of transcribing spontaneous speech, where transcribers tend to be guided by their expectations.

Whereas the phoneticians classified only 21.8% of all /t/s as absent, the ASR tool classified 31.7% of /t/s as absent. This may not come as a surprise, since the acoustic phone models were trained on read speech, where a more canonical realization of /t/ is more likely than in conversational speech. What is more, the acoustic models were trained at a frame shift of 5 ms and they consist of three emitting states. Hence, segments have a minimum length of 15 ms. In reality, [t]s may be shorter than that. Very short segments can be detected in the forced alignment, but at the cost of somewhat inaccurate segment boundary placement.

Another possible explanation for the discrepancy is that certain sub-segmental properties that are used by human listeners may be ignored by the ASR system. Whereas the presence of a constriction, a burst, and of alveolar friction were significant predictors for both, the MFCC parameters used by the ASR system make it difficult to discover whether the friction of the following segment starts smoothly or abruptly, even with the presence of delta and delta-delta coefficients. Furthermore, the ASR system annotates fewer [t]s as present if the burst is weak than if it is strong. Humans, in contrast, detect [t]s independently of the type of burst. Finally, there is a cue that is potentially misleading for the ASR system: Voiced realizations of /t/ are more likely to be annotated as absent than unvoiced ones. One might explain this result by arguing that the acoustic model for /t/ was trained on mainly voiceless realizations of /t/ because the training material was read speech, where canonical pronunciations are more frequent. However, since our set of pronunciation variants offered the ASR system the option to annotate voiced realizations as /d/, we draw the conclusion that voicing indeed is a misleading cue for the ASR system.

5. Summary and Conclusions

The aim of the present study was two-fold. The first aim was to investigate which of the sub-segmental properties of /t/ predict its perceptual presence for human listeners. We saw that the two phoneticians more often classify a /t/ as acoustically present if it contains a constriction, a burst, alveolar friction and if the friction of the following consonant starts abruptly. In the future, we hope to perform a similar study with naive listeners. These findings inform models of speech comprehension, which have to take sub-segmental variation into account in order to explain the processing of every day speech.

Secondly, we investigated which sub-segmental properties an ASR systems relies on and we compared the results with those from human listeners. Both are sensitive to the presence of a constriction, a burst and alveolar friction. However, there is a misleading cue for the ASR system, that is, voicing, and the system is less sensitive to fine cues (i.e. weak bursts, smoothly starting friction) than the human ear. These results suggest that automatic transcription tools can achieve better performances if they first identify the intervals where [t] may be present (given the canonical segmental representation of a word) and then apply detailed analysis and classification techniques informed by knowledge about how human perception is based on sub-segmental properties.

6. Acknowledgements

This research was supported by the Marie Curie Research Training Network 'Sound to Sense'. Mirjam Ernestus was supported by a European Young Investigator award from the European Science Foundation. The authors would like to thank Lou Boves for useful comments on an earlier version of this paper.

7. References

[1] B. Schuppler, W. van Dommelen, J. Koreman, and M. Ernestus, "Word-final [t]-deletion: An analysis on the segmental and sub-segmental level," in *Proceedings of Interspeech*, 2009, pp. 2275–2278.

[2] E. Janse, S. G. Nooteboom, and H. Quené, "Coping with gradient forms of /t/-delition and lexical ambiguity in spoken word recognition," *Language and Cognitive Process*, vol. 22, no. 2, pp. 161–200, 2007.

[3] O. Scharenborg and L. Boves, "Pronunciation variation modelling in a model of human word recognition," in *Proceedings of Workshop on Pronunciation Modeling and Lexicon Adaptation*, 2002, pp. 65–70.

[4] D. Norris, "Shortlist: a connectionist model of continuous speech recognition," *Cognition*, vol. 52, pp. 189–234, 1994.

[5] O. Scharenborg, "Using durational cues in a computational model of spoken-word recognition," in *Proceedings of Interspeech*, 2009, pp. 1675–1678.

[6] C. Van Bael, "Validation, automatic generation and use of broad phonetic transcriptions," Ph.D. dissertation, Radboud Universiteit Nijmegen, 2007.

[7] B. Schuppler, M. Ernestus, O. Scharenborg, and L. Boves, "An automatic method to analyze acoustic reduction in a corpus of conversational Dutch," Submitted.

[8] K. Kirchhoff, G. A. Fink, and G. Sagerer, "Combining acoustic and articulatory feature information for robust speech recognition," *Speech Communication*, vol. 37, pp. 303–319, 2002.

[9] O. Scharenborg, V. Wan, and R. K. Moore, "Towards capturing fine phonetic variation in speech using articulatory features," *Speech Communication - Special Issue on Intrinsic Speech Variation and Speech Recognition*, vol. 49, pp. 811–826, 2007.

[10] M. Wester, S. Greenberg, and S. Chang, "A Dutch treatment of an elitist approach to articulatory-acoustic feature classification," in *Proceedings of Eurospeech*, 2001, pp. 1729–1732.

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

[12] 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.

[13] A. Hämäläinen, M. Gubian, L. ten Bosch, and L. Boves, "Analysis of acoustic reduction using spectral similarity measures," *Journal of the Acoustical Society of America*, vol. 126, 2009.

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

[15] 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.

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

[17] C. Cucchiarini and D. BinnenPoorte, "Validation and improvement of automatic phonetic transcriptions," in *Proceedings of ISCLP*, Denver, USA, 2002, pp. 313–316.