Perception of lost contrast

Marion Jaeger
Unit for Language Structure: Man and Machine
at the University of Konstanz
Universitätsstrasse 1, D-78464 Konstanz, Germany
marion.jaeger@uni-konstanz.de

ABSTRACT

Psycholinguistic research in speech recognition and lexical representation has focused upon regressive Place assimilation across word-boundaries to explain how listeners use spectral information to disambiguate neutralized phonetic information contained in the acoustic signal. The favored paradigms are gating and phoneme identification experiments. This paper presents experimental and statistical data derived from a large German corpus of spontaneous speech that suggest listeners interpret segments with respect to their position within a syllable, the phonetic cues contained in the region of maximum spectral transition, as well as phonological constraints of the language.

1. INTRODUCTION

Regressive Place assimilation is a term used by phonologists to describe a post-lexical process in speech production that leads to a loss of contrast between two lexically distinct segments, thus creating lexical ambiguity in the acoustic signal [1]. Under normal circumstances listeners resolve this ambiguity by means of phonological, semantic, and syntactic information [2]. The current study, however, is only concerned with basic contexts of speech perception, namely with the perception of unfolding phonetic information within and across adjacent segments. The question of how listeners use the time-varying spectral information in spontaneous speech to identify speech sounds is addressed by comparing the results of a timed, force-choice phoneme identification task and a gating experiment. In both paradigms the same stimuli spliced from a carefully recorded and annotated database of spontaneous dialogues (The Kiel Corpus of Spontaneous Speech, 1995) were used. By using 'assimilated' [labial] and [dorsal], 'unassimilated' [coronal], and respective control items the results may provide some insight into the underlying lexical representation and how it influences speech perception. Finally, as this study deals with physical events, data from a comprehensive statistical analysis [3] of the database is also considered in the evaluation of the perceptual tasks.

2. THE DATABASE AND METHODS

The Kiel Corpus of Spontaneous Speech [4] is a large computer accessible database of approximately four hours of spontaneous dialogues. Spontaneous speech was elicited by asking participants to arrange meetings, not knowing the other person's schedule in advance. All dialogues were recorded in a sound-treated room to ensure high quality speech. Acoustic signals were carefully transcribed by phoneticians using both visual (speech waveform and spectrographic displays) and auditory information. The conversations were also annotated with respect to the canonical pronunciation. For this study the enumeration of assimilations both in the transcribed pronunciation and the canonical phonemic representations comprised 107 dialogues consisting of 1837 turns of approximately 1-3 minutes from 32 German speakers (19 males and 13 females) using a Standard German dialect but different speaking styles.

3. STATISTICAL ANALYSIS

Statistical analysis of the corpus yielded 1865 possible contexts where neutralization of a Place contrast between (plosive or nasal) stops across word boundaries might occur. In 1368 cases (73.4%) the target was a [coronal] and the trigger a [labial] or [dorsal] stop. In 401 cases (21.5%) the target was [labial] or [dorsal] and the trigger [coronal]. In only 96 cases was the target a [labial] and the trigger a [dorsal] stop or vice versa (Table 1). Only [coronal]#[noncoronal] stop sequences in the canonical representations were transcribed as 'neutralized'. Their frequency (8.9%), however, was low and limited to certain segmental and prosodic contexts. In particular, words with a high frequency of occurrence were affected. Additionally, these words were always unaccented and consisted of one syllable which in most cases (88%) ended in a nasal stop preceded by a schwa or a front vowel. The observed preponderance of [nasal] relative to [plosive] targets can be accounted for both by the frequency and salience of nasals. Thus, 3808 word-final [t]s and 8151 word-final [n]s were counted in the canonical files of the Kiel corpus. 36.6% of the word-final [t]s, but only 10.6% of the word final [n]s were transcribed as 'deleted' in the pronunciation files. The lower rate of [nasal] stop deletions could be due to the observation that even if a nasal is deleted, nasalization of the preceding vowel may still be a very robust perceptual cue for a nasal (even when deleted). The trigger in [coronal]#[noncoronal] stop sequences was, in most cases, voiced, suggesting that Place assimilation should not be viewed apart from other features. Regressive assimilation of [labial] and [dorsal] to a [coronal] target was equally frequent. The seven neutralizations of a [noncoronal] target (12%) could be accounted for by other reasons such as deletion of the schwa and the word-final [labial] prior to assimilation of a [dorsal] to a [coronal] nasal ([ainəm]#[glas]) or simply to deletion of the [əm] ending (e.g. [ainəm]#[tʰʀefn]).

ISBN 1-876346-48-5

Target	[coronal]		
Trigger	Context	Cano	Transc
[labial]	n_m	270	41
	n_b	308	46
	n_p	43	10
	t_m	233	16
	t_b	100	1
	t_p	43	1
[dorsal]	n_k	146	27
	n_g	88	15
	t_k	52	0
	t_g	85	1
Σ [coronal]#[noncoronal]		**1368**	**158**
Target	[noncoronal]		
Trigger	Context	Cano	Transc
[coronal]	ŋ_n	4	3
	ŋ_d	9	0
	ŋ_t	5	0
	k_n	22	0
	k_d	56	0
	k_t	30	0
	m_n	46	3
	m_d	120	0
	m_t	66	0
	p_n	4	0
	p_t	6	0
	p_d	33	0
[dorsal]	m_g	13	1
	m_k	29	0
	p_g	0	0
	p_k	1	0
[labial]	ŋ_m	8	0
	ŋ_b	3	0
	ŋ_p	0	0
	k_m	18	0
	k_b	23	0
	k_p	1	0
Σ [noncoronal]#[noncoronal]		**497**	**7**

Table 1. Regressive Place assimilation across word-boundaries in the *cano*nical and *transc*ription files

In sum, Place neutralizations is not very frequent in spontaneous speech. It is clearly asymmetric across word boundaries: only [coronal] stops are 'assimilated' by [noncoronal] stops but not vice versa. The preference of [coronal] targets seems to be due to the pronunciation variation observed in spontaneous speech. This notion is supported by the observation that nearly all [coronal] targets were nasals. Words, in which a word-final [coronal] stop undergoes a categorical sound change are highly frequent, unaccented and always contain a mid to low front (unrounded) vowel or a schwa. The trigger was usually a voiced [labial], an unvoiced [dorsal] plosive, or a nasal.

4. PHONEME IDENTIFICATION TASK

Prior to the gating experiment a timed, forced-choice phoneme identification task was run to insure that the assimilations carefully labeled by trained phoneticians under visual and auditory control were perceived as well by naive listeners. The selection of our stimuli was constrained by the stimuli found in the Kiel corpus and the following pragmatic concerns: (1) the segmental context should be as similar as possible, hence 'assimilated', 'unassimilated' and control utterances stem from different speakers; additionally, prevocal segments were discarded, and only vowel-stop sequences were presented (e.g. d[en]#können, 'then could') to keep the context identical, (2) the target segment had to be a nasal due to paucity of cases in which a word-final plosive was 'assimilated', (3) the vowels had to be a mid or low front vowel, since the schwa in 'assimilated' items was in most cases too short to be gated, and because 'assimilated' utterances containing a schwa yielded a greater variety, and hence a fewer number of stimuli to choose from. Utterances containing a front vowel in most cases started with [dan] ('then') or [den] (phonetic variant of [dan]). From these eight pairs of utterances stimuli were selected, each from a different speaker. Depending on the speaker the word-final nasal target in these utterances was either transcribed as 'unassimilated' or 'assimilated' (e.g. [da<u>n</u>]#[<u>b</u>is] versus [da<u>m</u>]#[<u>b</u>is] 'then till'). The trigger was equally divided between a voiced or unvoiced plosive. Control items were spliced from [dorsal]#[dorsal], [labial]#[labial], or [coronal]#[coronal] context (e.g. [dan]#[t^{h}ʀefn] 'then meet'; [kʀaːm]#[pasn] 'fits me'; [am]#[bestn] 'the best'). Two [dorsal] control items were not preceded by a [dorsal] plosive but by a pause. Unfortunately [eŋ] in German does not only constitute a syllable but also a word ('tight'). The splicing of all vowel-nasal sequences was carried out under auditory and visual control at zero crossings. Before recording the stimuli onto a DAT-tape their rms-amplitude was equated. Each token was preceded by a short warning tone. 44 naive listeners participated in the experiment. (Of the 49 subjects, five were excluded because their percent correct identifications and/or RT differed two standard deviations from the group means.) Listeners had to decide whether the nasal in a vowel-nasal sequence was [m], [n], or [ŋ] and press the appropriate button with their dominant index finger as quickly as possible. Reaction times (RT) was measured from the end of the signal and percent correct identifications were calculated. Two items had to be discarded from further analysis. One was perceived 'assimilated', the other 'unassimilated', but both had been transcribed as the opposite. Analyses of variance was performed, with the dependent variables being *percent correct score* and *mean RT* and the independent variables being *Place* and *condition within Place*. With respect to percent correct scores, control items differed by Place: [coronal] controls yielded a significantly lower percent correct score than [noncoronal] controls. Within conditions, 'unassimilated' [coronal]#[noncoronal] stimuli differed significantly from [coronal]#[coronal] control items, as did 'assimilated' and [noncoronal] control stimuli. Still, percent correct scores for 'assimilated' items was high: 82.5% of 'assimilated' segments were classified as [dorsal] or [labial] (Table 2).

ISBN 1-876346-48-5

Condi-tion	Context ***Place perceived***	Percent correct *Condition within Place*	Percent correct *Condition*
A	[coronal]#***[dorsal]***	81.2	82.5
A	[coronal]#***[labial]***	82.9	
C	***[dorsal]***#[dorsal]	92.5	94.3
C	***[labial]***#[labial]	96.2	
C	***[coronal]***#[coronal]	83.3	83.3
NA	***[coronal*** #[dorsal]	72.8	71.5
NA	***[coronal]***#[labial]	70.2	

Table 2: perception of 'assimilated' (A), 'unassimilated' (NA), and 'control' (C) items; word-boundary (#).

Results for mean RTs were in concordance with percent correct score: [coronal] controls were identified significantly more slowly (599 ms) than [dorsal] and [labial] controls (528 ms and 529 ms), and RTs of 'unassimilated' (576 ms) and 'assimilated' (583 ms) items differed significantly from their [coronal] and [noncoronal] control conditions. The difference between [coronal] and [noncoronal] control segments can be accounted for by their differential likelihood of deletion in word-final position. In the Kiel corpus [coronal] stops in word-final position are much more likely to be deleted or substituted than noncoronals [3]. This finding is in accordance with an analysis of the Switchboard corpus (Am. English) [5], in which deviations from the canonical form of coda segments in spontaneous speech was reported to be far less likely for [dorsal] and [labial] as compared to [coronal] consonants. The latter was found to be extremely likely to be non-canonically realized and thus harder to perceive. The even lower percent correct scores for 'unassimilated' [coronal] items are most likely due to coarticulatory cues from the following (although not presented) plosive. Earlier studies [6] showed that listeners involved in a phoneme identification task are sensitive to partial cues of neighboring segments and that subcategorical phonetic mismatch can interfere with listeners' phonetic judgments. In this line, the difference between 'assimilated' and [noncoronal] controls may be accounted for. Thus, in 'assimilated' segments with a residual coronal the vowel-nasal transition might have been perceived inappropriately due to an incomplete overlap and thereby imposed a processing load.
In sum, listeners process [coronal] and [noncoronal] coda segments differently. The difference in auditory perception can be accounted for by the tendency of coronals to be non-canonically realized in coda positions and listeners inclination to take into account all available cues, when making a phonetic judgment.

5. GATING EXPERIMENT

The gating paradigm provides a method by which to study perceptually relevant information along the temporal dimension [7,8]. In this experiment successive backward gating was used. Listeners heard increasing portions of the complete 'assimilated', 'unassimilated', and control vowel-nasal sequences and had to retrieve a disyllabic word starting with a vowel. To test the influence of phonological constraints on speech perception, the last gate included the burst of the trigger (e.g. d[anb]is, 'then till'). Truncated syllables are edited from the stored waveforms under auditory and visual control. To avoid clicks introduced by the truncations, all segments were spliced in the vicinity of zero crossings. An alignment point was set at the vowel offset in each vowel-nasal-plosive sequence. A sequence of gates continued from the alignment point in 10ms steps into the vowel and the stop portion, creating three vowel gates and five consonant gates. Due to speaker-dependent variability of the utterance, the first vowel and the last two consonant gates were of variable length. Before constructing the experimental tape the rms-amplitude of the tokens was equated. The tape consisted of at least three practice trials followed by 28 test items, four from each condition and context. The same items as in the phoneme identification task were used. 35 listeners took part in the gating experiment. Listeners responses to the stimuli were subsequently classified as either words or nonwords (single vowels, unknown or not in the dictionary). For each word choice the Place feature of the postvocalic consonant was judged as either correct or incorrect. For example, the feature [labial] in an 'assimilated' [coronal]#[labial] context was scored as correct but as incorrect in an 'unassimilated' [coronal]#[labial] context. In the latter case a word with a postvocalic [coronal] consonant would have been correct. First inspection of the items (Table 3) (only vowel-nasal gates) showed the same items had to be discarded as in the phoneme identification task. In addition, a third item was perceived significantly worse than the rest. Most listeners heard a rounded, mid front vowel and were unable to find a word, or if they perceived an unrounded mid front vowel, their response to the nasal at all gates was at chance level. For the rest of the items, listeners percent-correct scores improved with each gate (G1: 40.1%, G2: 44.6%, G3: 46%, G4: 55%, G5: 65%, G6: 70%, G7: 73%). The difference of correct scores between gate 2 to 3, 3 to 4, and 4 to 5 ($p<0.001$) was highly significant. These gates extend from -10ms before the alignment point to +20ms after. With respect to control and 'unassimilated' conditions, as well as control and 'assimilated' conditions, items differed significantly by overall correct responses, just as they did in the phoneme identification task. Listeners' overall percent-correct scores ranged between 43% and 71% (mean: 65%, SD: 7.9%). As none of the listeners performed more than two standard deviations below the mean, and all had understood the task, none were excluded from the analysis. Missing values ranged between 0.6% and 16%, indicating that the task was rather difficult.
With respect to gate 8, in which the trigger was presented, percent-correct scores for 'assimilated' and all control items improved slightly, whereas percent-correct scores for 'unassimilated' stimuli dropped. For example, a nasal perceived as [coronal] up to gate 7 was at gate 8 perceived as [labial] (e.g., the word choice *Anton*, 'Anton' becomes *Ampel* 'traffic light' after the burst of the [labial] plosive was presented). This change in perception is most likely due to phonological constraints. In German the sequence [n#p] is far less frequent than [m#p] and this might have biased listeners perception.

 ISBN 1-876346-48-5

Cond_Place	I	G1	G2	G3	G4	G5	G6	G7
*Ass*cor_dor	1	43	51	29	26	46	46	51
*Ass*cor_dor	2	31	29	31	43	51	80	74
*Ass*cor_dor	3	11	26	23	49	54	60	66
*Ass*cor_dor	4	49	63	66	69	69	74	71
*Ass*cor_lab	5	51	51	40	14	3	6	0
*Ass*cor_lab	6	29	34	46	31	43	63	49
*Ass*cor_lab	7	66	66	63	51	63	66	57
*Ass*cor_lab	8	37	40	57	69	80	94	94
*Ctrl*cor_cor	9	37	43	69	83	74	71	66
*Ctrl*cor_cor	10	60	43	40	63	69	80	86
*Ctrl*cor_cor	11	49	46	40	49	63	63	63
*Ctrl*cor_cor	12	43	54	54	57	69	80	80
*Ctrl*dor_dor	13	20	17	11	43	57	77	91
*Ctrl*dor_dor	14	14	9	20	34	54	77	89
*Ctrl*dor_dor	15	46	63	60	60	74	80	83
*Ctrl*dor_dor	16	54	71	77	83	91	97	97
*Ctrl*lab_lab	17	46	54	40	49	46	49	60
*Ctrl*lab_lab	18	54	71	69	66	86	80	91
*Ctrl*lab_lab	19	34	57	71	77	80	80	80
*Ctrl*lab_lab	20	54	60	57	69	83	86	89
*Non*cor_dor	21	43	37	37	49	69	60	63
*Non*cor_dor	22	40	40	49	43	46	46	31
*Non*cor_dor	23	31	31	40	46	66	57	77
*Non*cor_dor	24	54	51	40	60	74	80	83
*Non*cor_lab	25	29	31	51	51	51	43	51
*Non*cor_lab	26	31	26	40	34	51	54	60
*Non*cor_lab	27	40	43	29	43	43	46	60
*Non*cor_lab	28	26	31	34	51	54	66	46

Table 3: Percent correct scores by item (I) and gate (G); framed items were excluded from statistical analysis, 'unassimilated' (*non*), 'assimilated' (*ass*), control (*ctrl*), coronal (cor), labial (lab), dorsal (dor).

In summary, the results of the gating experiment show that in this task which is considered to tap the lexicon [2] listeners are sensitive to subcategorical phonetic cues in the same way as they are in a phoneme identification task. Second, the critical portion within the vowel-nasal sequence in which correct identification for the Place feature of the nasal increased significantly was located in the vowel-nasal transition region. Third, the drop in percent correct responses after the presentation of the onset of the second syllable suggests that listeners' perception is influenced by phonological constraints.

6. GENERAL DISCUSSION

Statistical and perceptual data have been presented that show that listeners' perception is influenced in many ways. Statistical data suggest that listeners variations in pronunciation of segments with respect to the position within the syllable influence how well they are perceived. Perceptual data are in accordance with earlier studies that showed that listeners are sensitive to partial cues of adjacent segments. Finally, listeners categorical identification seems to rely on a circumscribed region of the waveform in the vicinity of the vowel-nasal transition. This result is consistent with the results reported by Furui [7] who showed that performance on consonant recognition changed from 60% to 85% from a gate -10 ms before the critical point to the actual critical point. The critical point coincided with the regions of maximal spectral change between the consonant and vowel. Since the same region was critical for initial and final truncations he concluded that the crucial information for both is contained across the critical region.

ACKNOWLEDGEMENTS

I thank Steven Greenberg for his useful comments and suggestions, and Henning Reetz for help with the perceptual experiments. This study was supported by the Leibniz Prize of the Deutsche Forschungsgemeinschaft (DFG) to Prof. Dr. A. Lahiri.

REFERENCES

[1] U. Frauenfelder, A. Lahiri, "Understanding words and word recognition", In W. D. Marslen-Wilson (ed) *Lexical Representation and Process*. Cambridge: MA: MIT Press, 1989, pp. 319-341.

[2] W. D. Marslen-Wilson, A. Nix, G. Gaskell, "Phonological variation in lexical access: Abstractness, inference and English place assimilation". *Language and Cognitive Processes* **10**, 1995, pp. 285-308..

[3] M. Jaeger, "Lost contrast in naturally spoken utterances" (submitted).

[4] www.ipds.uni-kiel.de/forschung/kielcorpus.en.html.

[5] S. Greenberg, H. M. Carvey, L. Hitchcock, S. Chang, "Beyond the phoneme – A juncture-accent model for spoken language". *Proceedings of the Human Language Technology Conference, San Diego, California,* 2002.

[6] D. Whalen, "Subcategorical phonetic mismatch slow phonetic judgments". *Perception and Psychophysics* **35***,* 1984, pp. 49-64.

[7] S. Furui, "On the role of spectral transition for speech perception". *J. Acoust. Soc. Am.* **80 (4)**, 1986, pp. 1016-1025.

[8] R. Smits, N. Warner, J. McQueen, A. Cutler, "Unfolding of phonetic information over time: A database of Dutch diphone perception". *J. Acoust. Soc. Am.* **113 (1)**, 2003, pp. 563-574, 2003.