Using acoustic cues in stop perception

Rolf Carlson
KTH, CSC, Dept. Speech, Music and Hearing

Abstract

In this paper we will return to a basic question in phonetics: the relation between acoustic cues and perceived phonetic units. One focus is to study how multiple cues distributed in time unite to form a probable perceptual identity. Using mixed waveforms we evaluate the perceptual impact of acoustical cues in unvoiced stops. The result shows that cues in a vowel preceding a stop can have a significant influence on the percept if the cues are in concordance with the stop release cue. This is the case even if the vowel cues are not strong enough to change stop identity by themselves. The second focus is to examine how the context actually can reduce the perceptual importance of stop release cues. By the use of repeated releases we show that a temporal distortion can have the same kind of impact on a percept as a spectral distortion. If the manipulation is classified as a disturbance, the cues have reduced importance for the classification.

Introduction

In this paper we will return to a basic question in phonetics: the relation between acoustic cues and perceived phonetic units. The research has been focused on the perception of unvoiced stops. A straightforward waveform splicing technique has been used to create stimuli with sometimes contradicting acoustical cues. The rationale for conducting the experiments is to further illuminate and support the hypothesis that speech perception is a dynamic and adaptive perceptual process in the interpretation of acoustic cues.

A special inspiration for the research has been the increased interest in fine phonetic details and the studies of temporal integration in speech perception (Hawkins, 2003). An unpublished pilot experiment, carried out during the 70s by Gunnar Fant, gave an intriguing illustration of the active processing in speech perception. The third formant in a natural front vowel was moved with the help of a pole-zero filter resulting in a perceptual vowel shift. However, if a sequence of different vowels were filtered with this stationary setup the perceived vowel identity was not shifted. The perceptual process was able to identify the filtering as a distortion and disregard that a formant was misplaced. A similar experiment was done by Eriksson et al. (2003) with the motivation to make speech more nasal, but again, without significant success. Experience from work on synthesis is further supporting that stationary processing does not solve problems with speech quality or phonetic identity. In the experiments reported here we will try to change the percept of a stop consonant and again with the help of a temporal context change it back to the original identity.

Background on stop features

Already the classical experiments on the perception of stops (Liberman et al., 1952) are looking for universal cues that can be used as robust features in speech perception. The results are proposed to point to an invariance that should be search for on an articulatory level rather than an acoustic one. However, the persistent work by Stevens on invariant cues points to how acoustic landmarks and features are used in human perception, see for example Stevens (2002). A review of speech perception including several studies on stops has recently been published by Hawkins (2004).

The acoustic modelling of the production of stops has been well described in the classical publications by Fant (1960, 1969), Halle et al. (1957) and Stevens (1998). However, a unifying model of stop coarticulation based exclusively on acoustic analysis is not easy to formulate. The work by Hawkins and Slater (1994) shows how coarticulation in terms of locus equations depends on a multitude of factors. Studies of stop cues, locus equations and the relation to articulation have been discussed in detail by Lindblom (1998) and Diehl (1998).

In the perceptual experiments by Schatz (1954) and Carlson et al. (1972) the search for acoustical cues is carried out by replacing segments, sometimes even filtered ones, in speech waveforms. The perceptual importance of stop release cues is clearly manifested. The results from these studies also have bearing on current datadriven synthesis models proposed by Carlson et al. (2002) and Hertz (2002), which mix parametric synthesis with naturally spoken segments.

Experiment

In the current experiment we study how multiple cues unite to form a probable perceptual identity. Using three classical manipulations we form a baseline for the fourth type of stimuli, where we evaluate if context can reduce the perceptual impact of stop release acoustic cues.

Stimuli preparation

Clearly spoken nonsense words by one speaker were selected as ***original*** stimuli. They had earlier been recorded (16 kHz sampling rate) for the development of a datadriven speech synthesis system (Carlson et al., 2002; Carlson and Granström, 2005). Eighteen nonsense words /te_'C_V_de/ were used including one of three unvoiced stop consonants (p,t,k) before one of six vowels (a, a:, i, i:, u, u:). Stress was on the second syllable. Thus, in total 18 stimuli were included in the stimulus group type *original*.

In the stimulus type ***initial,*** the first syllable in each original word /te_C_V_de/ was replaced by the corresponding part of another stimulus word. The inserted segment came from a word with a different consonant C but the same vowel V. The mixing point was placed in the stop gap of the consonant C. In total 36 stimuli were included in the stimulus group type *initial*.

In the ***release*** type of stimulus only the stop release in C (40 ms) in a word was replaced by another equally long stop release. The new segment came from a word with a different consonant C but the same vowel V. A few samples before and after the mixing points were interpolated to avoid an unwanted distortion. In total 36 stimuli were included in the stimulus group type *release*.

The combination of the two last types form the ***initial+release*** type replacing the first part of a word with another corresponding part. As a result only part of the aspiration of the original consonant C and the two final syllables are kept intact from the original word, see Figure 1. In total 36 stimuli were included in the stimulus group type *initial+release*.

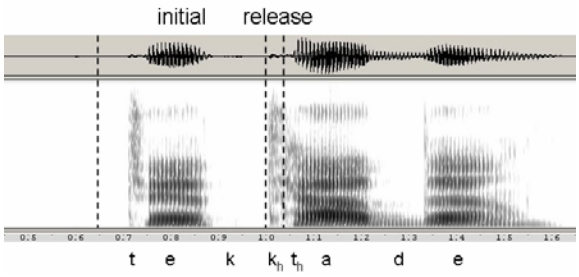

*Figure 1. Example of a type **initial+release** stimulus. The first part, the initial and the stop release (40ms), from "te'kadde" has replaced the corresponding part in "te'tadde".*

Finally, the ***repeated release*** type stimuli has the same processing as the *release* type but in addition the replacing release in consonant C is repeated at regular intervals. The sequence of repeated releases creates a distortion in the stimuli. It is important to remember that one replacing release is still at the same position in consonant C as in the *release* type, see Figure 2. In total 36 stimuli were included in the stimulus group type *repeated release*.

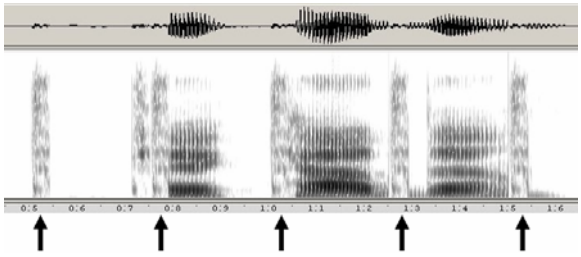

*Figure 2. Example of a repeated release stimulus. The release (40ms) in "te'**k**adde" has replaced the release in "te'**t**adde" (middle arrow). Furthermore, the k-release is repeated at regular intervals (marked with arrows).*

Subjects and experiment procedure

Seven subjects working at the department participated in the experiment. Not aware of the hypotheses and motivations for the experiment, they were instructed to simply report which unvoiced stop (p, t, k) they heard in the middle of the respective nonsense word. Perceptual data for the individually randomized sequence of 162 stimuli was collected using a web interface. If needed, it was possible for the subject to repeat a stimulus. The experiment was typically carried out within half an hour for each subject.

Results

We will in the following use the abbreviation **COP** (change of percept) when we discuss the perceptual results. An example of such a change is when a p-release replacing a t-release in /te'tade/ changes the perception of the word to /te'pade/. Of the 144 manipulated stimuli, the subjects changed the identity of the consonant C in 69 cases as an average with a standard deviation of 6. The COP results grouped according to stimulus type and vowel length are presented in Figure 3.

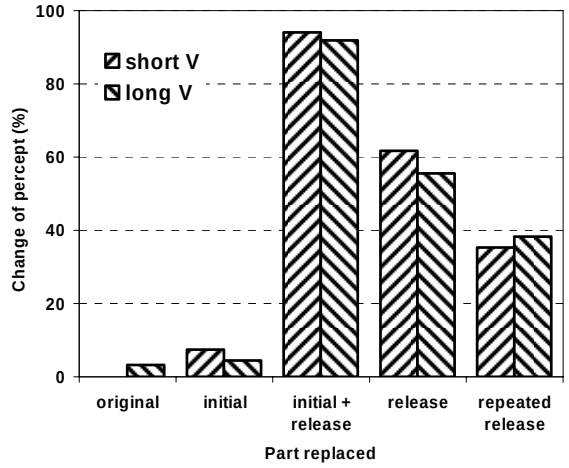

Figure 3. Change of percept (COP) grouped according to stimulus type and vowel length.

As expected the *initial+release* type has a very high COP (93%), while the *initial* type has very little impact on the consonant identity (6%). More than half of the stimuli changed their identity for the *release* type (59%), where only the 40 ms burst segment was replaced.

COP has no significant correlation with either the voice onset time (VOT) for the three stops or the following vowel duration, Figure 4. On the contrary, the original /t/ stimulus gets higher COP than both the /p/ and /k/ stimuli, while the VOT for /t/ is shorter than for /k/ and longer than for /p/, see Figure 5. Neither have the phonological vowel length or the original stop gap duration significant impact on the COP.

Finally, the *repeated release* type has a COP of only 37% compared to the 59 % COP for the *release* type.

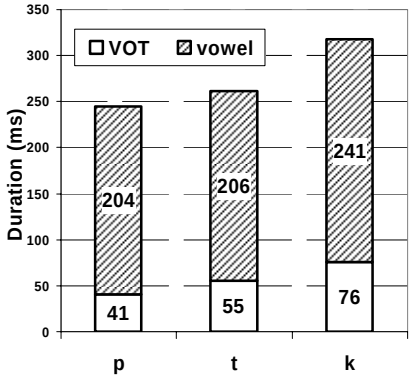

Figure 4. Average duration of voice onset time (VOT) and following vowel in the original stimuli.

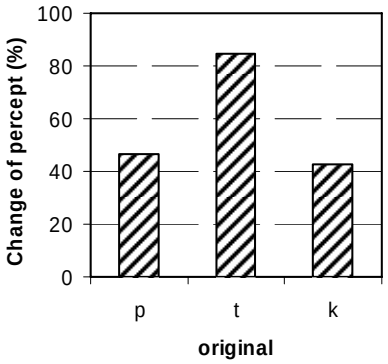

*Figure 5. Change of percept (COP) of **release** type grouped according to **original** consonant.*

Discussion

The COP results for the *initial* type and the *release* type are according to what can be predicted based on general knowledge of speech perception and the publications reviewed in the introduction. The cues in the proceeding vowel are weaker than the cues in the stop release. However, it is interesting to note that the combination of acoustical cues in the proceeding vowel and the stop release generates a stronger COP than a simple addition of the results from the individual sets of cues. One could then speculate that the cues in the proceeding vowel play an important secondary role by adding robustness to the percept. A coherence of the acoustic cues makes the decision less confusing for the listener.

The strong COP dependence on the original stop release identity is worth noting, Figure 5. Furthermore, the replacement of the acoustic cues in the /t/ release is much more harmful than for the other stops. However, the reverse is not valid. Insertion of a /t/ release does not change the COP more than the /p/ or /k/ release does.

How the subjects perceived the *repeated release* type stimuli compared to the *release* type is the other focus of the experiment. By repeating a sequence of releases we introduce an unnatural distortion. The COP results suggest that listeners correctly classified the repeated releases as a distortion and thus tried to disregard the disturbing acoustic releases in the identification process. Unfortunately for the listener this also applied to the correctly aligned stop release. Thus, the replacing stop release cues received less importance compared to the same replacement in the release type of stimuli.

The study was based on one set of nonsense stimuli pronounced by one speaker. The work needs to be confirmed by additional experiments using varying conditions and more speakers.

Conclusion

By the use of repeated releases we show that a temporal distortion can have the same kind of impact on a percept as a spectral distortion. If the manipulation is classified as a disturbance, the cues have reduced importance for the classification. Furthermore, the result shows that acoustic cues in a vowel preceding a stop can have a significant influence on the percept if they are in accordance with the stop release cues.

Acknowledgements

We thank the subjects participating in the experiment.

References

Carlson R & Granström B (2005). Data-driven multimodal synthesis. *Speech Communication*, 47(1-2), 182-193.

Carlson R, Granström B & Pauli S (1972). Perceptive evaluation of segmental cues. *In Proceedings of the Conference on Speech Communication and Processing* (pp. 206-209). Bedford, MA, USA. also STL-QPSR, 13(1), 018-024.

Carlson R, Sigvardson T & Sjölander A (2002). Data-driven formant synthesis. *Proceedings of Fonetik*, TMH-QPSR, 44(1), 121-124.

Diehl R L (1998). Locus equations: A partial solution to the problem of consonant place perception. *Behavioral and Brain Sciences*, 21, 264.

Eriksson E, van Doorn J, Sullivan K (2003). Turning wine into water: Can ordinary speech be artificially nasalized? *Umeå University, Department of Philosophy and Linguistics PHONUM* 9, 145-148

Fant G (1960). *Acoustic Theory of Speech* Production Mouton De Gruyter; Revised edition (January 1970)

Fant G (1969). Stops in CV-syllables. *STL-QPSR* 4/1969, pp. 1-25.

Halle M , Hughes G , Radley J-P (1957). Acoustic properties of stop consonants", *Journal of the Acoustical Society of America* 1-29, 107.

Hawkins S (2003). Roles and representations of systematic fine phonetic detail in speech understanding. *Journal of Phonetics* 31: 373-405.

Hawkins S (2004). Puzzles and patterns in 50 years of research on speech perception. *From Sound to Sense: 50+ Years of Discoveries in Speech Communication*, MIT, Cambridge, MA, USA

Hawkins S & Slater A (1994). Spread of CV and V-to-V coarticulation in British English: Implications for the intelligibility of synthetic speech. *ICSLP 94 (Proceedings of the 1994 International Conference on Spoken Language Processing)*, 1: 57–60.

Hertz S (2002). Integration of rule-based formant synthesis and waveform concatenation: a hybrid approach to text-to-speech synthesis. *Proceedings IEEE 2002 Workshop On Speech Synthesis*, Santa Monica, CA.

Liberman A, Delattre P & Cooper F S (1952). The role of selected stimulus-variables in the perception of unvoiced stop consonants, *Am. J. of Psych.*, 497-516.

Lindblom B (1998). An articulatory perspective on the 'locus equation'; commentary on target article by Sussman, Fruchter, Hilbert, Sirosh, Linear correlates in the speech signal: the orderly output constraint. *Behav. Brain Sci*. 21: 241–299.

Nguyen N & Hawkins S (2003). Temporal integration in the perception of speech: introduction. *Journal of Phonetics*, 31, 279-287.

Schatz C (1954). The Role of Context in the Perception of Stops. *Language*, Vol. 30, No. 1: 47-56

Stevens K N (1998). *Acoustic phonetics*. Cambridge, MA/London: MIT Press.

Stevens K N (2002). Toward a model for lexical access based on acoustic landmarks and distinctive features. *Journal of the Acoustical Society of America*, 111, 1872–1891.