WHEN IS FINE PHONETIC DETAIL A DETAIL?

Rolf Carlson and Sarah Hawkins

KTH, CSC, Dept. Speech, Music and Hearing, University of Cambridge
rolf@speech.kth.se, sh110@cam.ac.uk

1. INTRODUCTION

It is our task to take a discussant role in the special session "Sound to Sense: Modelling Fine Phonetic Detail" at ICPhS 2007. The contributions by Moore and Maier [12] and Lecumberri and Cooke [11] inspire further thinking on how fine phonetic details can be successfully explored by humans or by machines. The MINERVA2 system built on the multi-trace (episodic) memory model challenges current traditional probabilistic efforts in speech recognition by including a more human-like approach. The rationale for the comments in this paper is to illuminate and support the hypothesis that speech perception is a dynamic and adaptive perceptual process in the interpretation of acoustic cues or fine phonetic details. As background for the discussion of the two contributions two experiments are reviewed.

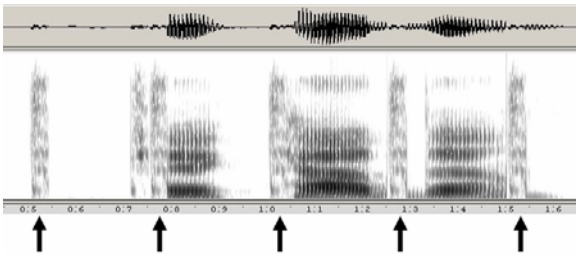
2. BACKGROUND: TWO EXPERIMENTS

An unpublished pilot experiment, carried out during the 70s by Gunnar Fant, gave an intriguing illustration of active perceptual processing of speech. The third formant in a natural front vowel was moved with the help of a pole-zero filter, resulting in a perceived 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: for the listener, F3 was not displaced, but indicated something else, perhaps speaker identity. This effect complements the finding that the perceived vowel quality—and lexical identity—of a syllable can be influenced by the average F1 frequency of a precursor phrase [9, 10]. The particular acoustic manipulation, and its distribution in the signal, determines how it is responded to: whether as a distortion, which is a transmission (channel) property, or a speaker characteristic. Although speaker characteristics might be modelled as transmission channel properties in some circumstances, e.g. in multi-speaker conversations, we keep the two concepts apart conceptually.

In a second study, Carlson [3] tried to change the percept of a voiceless stop consonant, and then change it back to its original identity by changing its context. Parts of stops were spliced to make stimuli with contradictory acoustic cues. Three types of manipulation provided a baseline for the fourth type, which tested whether context could reduce the perceptual effect of stop release cues.

18 nonsense words /te'CVde/ were spoken by a Swedish speaker. C was one of three voiceless stops /p t k/; V was one of six vowels, /a a: i i: u u:/. From each original, 4 further stimuli were made. (1) In ***initial*** stimuli, the first syllable, /te/, in the original was replaced by the corresponding part of another stimulus. The inserted segment came from a word with a different consonant C but the same vowel V. The splice point was in the silence corresponding to the stop closure, just before the stop release. As a result the duration of the stop closure was also changed according to the inserted segment. (2) In ***release*** stimuli, 40 ms from the C burst was replaced by the equivalent portion of another C stop. (3) The combination of types (1) and (2) formed the ***initial+release*** type, replacing the closure and VOT of the CV syllable with the corresponding part from another syllable.

Figure 1: Example of a repeated release stimulus. The C release (40ms) in "te'tade" has replaced the release in "te'kade" (middle arow). Furthermore, the t-release is repeated at regular intervals (marked with arrows).



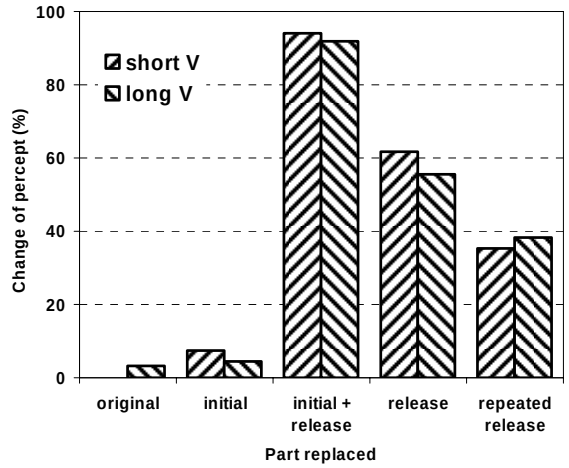
Finally, (4), ***repeated release*** stimuli were the same as type 2 (*release*) except that in addition the replacing release in C was repeated at regular

intervals throughout the stimulus. The sequence of repeated releases created a distortion in the stimuli. It is important to remember that one replacing release was still at the same position, in consonant C, as in the *release* type, as shown in Fig. 1.

Using a web interface, seven listeners, working at KTH CSC but naïve about this experiment, heard the randomized sequence of 162 stimuli (18 + 4(36)). They reported which voiceless stop /p t k/ they heard in the middle of each nonsense word. They could repeat a stimulus if needed.

The change of percept (COP) results grouped by stimulus type and vowel length are shown in Fig. 2. 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/. As expected, *initial+release* type has a very high COP (93%), while the *initial* type has little effect on consonant identity (6%). More than half (59%) the *release* type stimuli changed their identity.

Figure 2: Change of percept (COP) grouped according to stimulus type and vowel length.



Crucially, *repeated* release stimuli have a COP of only 37%, compared to 59% COP in *release* type. A logistic regression analysis (with COP as dependent variable, type as independent variable) showed release type responses were significantly different from all other types ($p<0.01$).

3. DISCUSSION

3.1. Context and signal

These data illustrate how humans combine multiple cues to form perceptual hypotheses, as reflected in sound identification. Two aspects are especially pertinent to the papers under discussion. First, the COP results suggest that listeners correctly classified the *repeated* releases as a distortion and thus tried to ignore the disturbing acoustic segments during identification. Unfortunately for the listener this also applied to the correctly-aligned stop release. Thus, the replacing stop release cues received less perceptual weight than the identical replacement manipulation in the *release* type of stimuli.

Here we have an example of world knowledge dictating the perceptual salience of an attribute of the signal in different ways depending on circumstances. The system needed to do this type of processing requires not just world knowledge, but also integration of information over some considerable time. This type of integration is a well-known attribute of auditory processing cf. [2], but poses a real challenge for computational modelling of speech recognition by humans and machines. One of the aims of S2S is to develop this type of long-domain temporal model.

In addition to demonstrating the fundamental nature of long-domain properties of perceptual decisions, these data also underline that perceptual decisions must be context-sensitive representations of certainty. Determining what is context, and what is signal, is presumably partly inherent to the perceptual system, and partly a function of individual experience, current expectation, and attention. In other words, perceptual decisions about speech will be very sensitive to signal properties and task demands. Ogden's discussion of the other papers develops this point in the more complicated area of conversation.

Second, the present experiment shows that, while cues in the preceding vowel are weaker than cues in the release, nevertheless the combination of cues in the preceding vowel and the release (*initial+release*) generates a stronger COP than a simple addition of the separate *initial* and *release* cues (93%>(6+59)%). The implication is that cues in the preceding vowel add robustness to the percept: when even weak cues are coherent, perceptual decisions are more consistent. This again underlines the importance of long-domain integration of acoustic information, and the central role of perceived context in this process.

Lecumberri and Cooke [11] describe how native listeners gained significantly more when contextual information in the preceding word was present, indicating that cross-word, extra-syllabic, cues are less easily exploited by non-native than by

native listeners. (Presumably this would not be the case if patterns of cross-word coarticulation were more similar in the two languages than they are in Spanish and English.) When the languages differ in coarticulatory patterns, a non-native speaker might need more exposure to learn how to combine several sources of language-dependent cues. There is a parallel in classical parametric speech synthesis, where only a limited number of obvious cues were modeled, resulting in less robust speech than could have been achieved with supportive secondary cues, as measured by intelligibility in adverse listening conditions, cf. [6, 15].

This type of issue is not restricted to non-native speakers. Lecumberri & Cooke observe that "some of the information needed to support a native advantage [was] removed by splicing" stimuli. Presumably this effect mainly reflects disrupted rhythm and f0—reduced continuity. Such data add to the classic literature that shows the intelligibility of speech fragments is strongly affected by hearing sufficient context e.g [14] and to more recent evidence that context is crucial when the speech is very casual [5]. The data of [11] show, first, that in noise, a small amount of context can significantly influence intelligibility even of carefully-spoken plosives preceding a single vowel, and second, that native familiarity with general speech patterns is necessary for benefits of context to be measurable.

Are the context-sensitive processes required to understand reduced speech different from the kind required to decide whether an unexpected sound like a stop release is relevant to the current speech signal? One set of data speaking to this point concerns so-called /r/ resonances, widely discussed extensively in the literature on fine phonetic detail. These /r/ resonances are acoustic reflections of an [ɹ] that may extend several syllables away from where the acoustic segment [ɹ] is identified [4, 7]. Some can be heard with careful listening—they were first documented through listening [8]—but it is our impression that most are not noticed in good listening conditions, although, to our knowledge, there have been no formal discrimination tests. Nevertheless, West [18] showed they are salient in natural speech when the /r/ segment is replaced by noise, while Hawkins and Slater [6] showed that their presence in synthetic speech can increase intelligibility in cafeteria noise by around 15%.

What is the perceptual basis of /r/ resonances? Are they another aspect of the type of continuity that Lecumberri and Cooke report, and that can perhaps be put under the general rubric of vowel-to-vowel coarticulation? In that case, they are likely to be quite language-specific, or knowledge-driven [1]. Yet, to what extent is that type of continuity a basic property of auditory processing, connected perhaps with auditory grouping cf. [16]? In other words, are /r/ resonances an unimportant detail, or a reliable, even fundamental, base? When is a fine phonetic detail a detail? It might be the perceptual glue upon which everything else depends, at least in adverse listening conditions.

3.2. Similarity representation in MINERVA2

In MINERVA2 all traces seem to have the same influence on the echo irrespective of their function. The fine phonetic details actually are more like fine acoustic details irrespective of their phonetic function. The current model does not yet seem to include a good technique to rank how much impact different acoustic details should have on the final perceptual outcome, as the experiments described above show. Invariant details due to stable acoustic properties discussed by Stevens [17] might automatically emphasize linguistically relevant acoustic-phonetic landmarks. This would enhance similarities in a group of traces of, e.g., a word.

In speech perception, parameters are often best represented on a logarithmic scale. This includes for example energy estimates, duration, intonation, frequency and spectral slope. One would like to see this reflected in the similarity measures used with MINERVA2. The similarity parameter in the model varies between 0 (no similarity) and 1 (full similarity). Thus, the similarity estimate needs to be raised to the power of p to facilitate non-linear behavior. Furthermore, this approach requires the similarity parameter to be normalized to be always between 0 and 1. This processing might introduce unnecessary complications in the model: the high p factor is actually slightly surprising in the final weighting model. A more intuitive approach might be that the similarity measure has the "top score" of 0 (full similarity) instead of 1. Then, the ad hoc normalization by the maximum distance measure to keep the similarity less than 1 could be replaced by a more stable normalization by the standard deviation in a z-score fashion. Such processing would be less dependent on the available training material and easily lend itself to logarithmic representation.

The second experiment in our section 2 deals with a time-dependent distortion reducing the

impact of the fine phonetic details in the stop release. This is a rather extreme manipulation and rarely found in real life. However, it is a challenge to build models for speech perception and speech recognition that offer a framework for seamlessly including new sources of information. Human perception is a dynamic adaptive process and future models need to handle such behavior. MINERVA2 has good potential in that respect. Perhaps this will be the point when human-inspired models outperform current probabilistic models.

The choice of test material will be important here. Moore & Maier justify their use of letters of the alphabet in terms of its small size and highly confusable items: there is almost no variation compared even with isolated monosyllables of English. As a test of a particular model, they have probably chosen an especially challenging data set.

But as a test of what parameters need to be included in a more general ASR model, we ask whether progress would be faster with a more linguistically diverse dataset. What insights could be gained from a corpus comprising short phrases that contrasted in carefully chosen morphological and grammatical distinctions, as well as phonemic ones, together with linked syntactic and prosodic trees, cf. [13]? Would the richer linguistic structure reduce the burden on the recogniser by changing the focus from identification of purely phonological units to identification of meaning available "in" the acoustic signal? Would emphasis on longer-term congruence be worth the extra modeling complexity?

These are not easy questions to answer. They illustrate the interesting problem of finding a good balance between tractable data and useful applications, but they go further: by simplifying the *data* and the *goals* of a recognizer, you may not develop the best *factors* in a model. These issues are central to the modelling aims of S2S.

4. CONCLUDING REMARKS

We have tried to show how the topics addressed by [11, 12] raise many of the most fundamental issues in developing better models of human and machine speech recognition. Further, we have tried to make clear that FPD is not all about tiny details, and is not all fine. The term FPD grew from a need to distinguish it from the standard "relevant to phoneme identification in citation-form words" assumptions. Some people think we need a new term that avoids the ragbag way that FPD is used today. They are right. The new term may simply be "phonetic information". It can replace FPD when standard models treat contextualised phonetic information as the norm. S2S aspires to building such models.

5. REFERENCES

[1] Beddor, P.S., Harnsberger, J., Lindemann, S. 2002. Language-specific patterns of vowel-to-vowel coarticulation: acoustic structures and their perceptual correlates. *J. Phonetics* 30, 591-627.

[2] Bregman, A.S. 1990. *Auditory Scene Analysis: The Perceptual Organization of Sound.* Cambridge MA: MIT.

[3] Carlson, R. 2007. Using acoustic cues in stop perception. *Proc. Fonetik 2007. TMH-QPSR* Stockholm, 50, 25-28.

[4] Coleman, J.S. 2003. Discovering the acoustic correlates of phonological contrasts. *J. Phonetics* 31, 351-372.

[5] Ernestus, M., Baayen, H., Schreuder, R. 2002. The recognition of reduced word forms. *Brain and Language* 81, 162-173.

[6] 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*, I, 57-60.

[7] Heid, S., Hawkins, S. 2000. An acoustical study of long domain /r/ and /l/ coarticulation. *Speech Production: Models and Data* Munich, 77-80.

[8] Kelly, J., Local, J.K. 1986. Long-domain resonance patterns in English. *Int. Conf. Speech Input/Output; Techniques and Applications* London, 258, 304-309.

[9] Ladefoged, P. 1987. A note on "Information conveyed by vowels". *J. Acoust. Soc. Am.* 85, 2223-2224.

[10] Ladefoged, P., Broadbent, D.E. 1957. Information conveyed by vowels. *J. Acoust. Soc. Am.* 29, 98-104.

[11] Lecumberri, M.L., Cooke, M.P. 2007. Effect of cross-word context on plosive identification in noise for native and non-native listeners. *Proc. 16th ICPhS* Saarbrücken.

[12] Moore, R.K., Maier, V. 2007. Preserving fine phonetic detail using episodic memory: Automatic Speech Recognition using MINERVA2. *Proc. 16th ICPhS* Saarbrücken.

[13] Ogden, R.A., Hawkins, S., House, J., Huckvale, M., Local, J.K., Carter, P., Dankovicová, J., Heid, S. 2000. ProSynth: An integrated prosodic approach to device-independent, natural-sounding speech synthesis. *Computer Speech and Language* 14, 177-210.

[14] Pickett, J.M., Pollack, I. 1963. Intelligibility of excerpts from fluent speech: Effects of rate of utterance and duration of excerpt. *Language and Speech* 6, 151-164.

[15] Pisoni, D.B. 1997. Perception of synthetic speech. In: van Santen, J.P.H., Sproat, R.W., Olive, J.P., Hirschberg, J. (eds), *Progress in Speech Synthesis*. New York: Springer. 541-560.

[16] Remez, R.E. 2003. Establishing and maintaining perceptual coherence: Unimodal and multimodal evidence. *J. Phonetics* 31, 293-304.

[17] Stevens, K.N. 2002. Toward a model for lexical access based on acoustic landmarks and distinctive features. *J. Acoust. Soc. Am.* 111, 1872-1891.

[18] West, P. 1999. The extent of coarticulation of English liquids: An acoustic and articulatory study. *Proc. 14th ICPhS* Berkeley, 3, 1901-1904.