SOUND TO SENSE: INTRODUCTION TO THE SPECIAL SESSION

Sarah Hawkins and John Local

University of Cambridge; University of York
sh110@cam.ac.uk, lang4@york.ac.uk

1. BACKGROUND

Sound to Sense (S2S) is a Marie Curie Research Training Network, funded 2007-2011. It involves some 50 researchers in 13 institutions in 10 countries. The ultimate aim of S2S is to provide models of speech processing that closely reflect the exquisite flexibility and robustness of human speech processing (HSP), that pave the way for the next generation of robust automatic speech recognition (ASR) and text-to-speech (TTS) machines, and that promise a new theoretical basis for foreign language (FL) teaching, and diagnosis and treatment of speech disorders. The immediate aim is to elucidate the interaction of knowledge and sensation in speech perception, using insights from recent linguistic, phonetic and psychological research to inform speech recognition models (HSP and ASR). The new models should better reflect the way humans listen and respond to their native language (L1) and to FLs. One focus is to track how phonetic information, especially fine phonetic detail that varies systematically with linguistic and interactional structure and function, is used in differing situations: when listeners have appropriate linguistic-phonetic knowledge (listening to L1), inadequate or inappropriate knowledge (listening to FL), and inadequate access to the signal (listening in adverse conditions). A related focus is to elucidate how phonetic information contributes to understanding.

S2S aims both to combine knowledge from independent disciplines and to reduce fragmentation within disciplines. Young research workers, with multidisciplinary training focused on a new theoretical framework, will acquire the perspective and skills needed to allow them to take speech processing research and applications further than any one discipline can currently achieve.

2. THEORETICAL MOTIVATIONS

Evidence from two independent strands of research converges to suggest that what humans do when they listen to speech may be radically different from the approach of the dominant HSP and ASR models. Dominant models typically assume that initial processing of speech involves transforming it into an abstract representation of discrete features or phonemes before further processing. But these two research strands, on episodic memory and on systematic variation in fine phonetic detail, suggest both that details of individual perceptions are remembered, and that, viewed in a richly-structured linguistic model [8, 13], much so-called phonetic variation varies systematically with the linguistic function of the stretch of speech, and can be perceptually salient [1, 7, 16, 17]. Such different assumptions have far-reaching consequences, promising to open up a more comprehensive and realistic theory of human speech communication than has been possible before, with consequent gains for technology and other applications. Inter- and intra-disciplinary divides have prevented fast progress, but the speech community now includes a critical mass of researchers with the motivation, knowledge, and skills to make concerted efforts to overcome them.

3. FINE PHONETIC DETAIL

The term ‘fine phonetic detail’ (FPD) was introduced some 20 years ago by John Local and colleagues to describe phonetic phenomena such as resonances associated with liquid consonants in English that were systematically distributed but not systematically treated in conventional approaches. Since then, the term FPD has been applied to anything that is not considered a major, usually local, perceptual cue for phonemic contrasts in the citation forms of lexical items. Experiments show that some FPD is indeed ‘fine’, and subtle, but other types are perfectly audible; they have just not been factored into the prevailing theory that perceptual processing of phonetic information is entirely aimed at identifying strings of features or phonemes that allow words to be distinguished. When this view is replaced, the term fine phonetic detail can simply be replaced, once more, by phonetic information.

Crucially, FPD does not just distinguish words, but also the wider phonological and

grammatical structure of the message. For example, grammatical function words have a narrower range of sound patterns than content words, and undergo different connected speech processes; and each type of function word (e.g. auxiliary verbs, articles) has its own distinct system of contrasts. FPD also reflects function and structure of the smaller units that comprise words, and of larger groupings, influencing everything necessary for successful communication: phonological, morphological, grammatical, pragmatic, interactional. FPD indicating a single linguistic distinction can involve many acoustic properties distributed over long stretches of speech [8, 13] e.g. traces of English /r/ can occur several syllables before the main /r/ segment and influence perception [2, 7, 11, 21, 22]. Thus, much FPD—the sort discarded by traditional abstractionist models as uninteresting or due to random effects—in fact systematically reflects many different aspects of meaning that are crucial to the maintenance of normal conversation: lexical, grammatical, and interactional differences. Even well-researched distinctions like coda voicing involve multiple distinctions, some of which are less local than was until recently assumed [9].

Not all FPD is perceptually salient in all conditions. This is intriguing, because FPD that is hard to detect in quiet can increase intelligibility in noise [7, 10]. Ignorance of FPD may explain the disproportionate difficulty of understanding an FL in noise [5]. For models to use FPD, we need to determine what types of FPD influence speech processing, under what conditions, and why.

That FPD can influence perception casts a new light on the old debate about the relative importance of top-down *vs* bottom-up information: instead of top-down information compensating for signal inadequacies, many S2S partners take the view that the signal is not interpretable in isolation from knowledge, and that the signal itself can indicate what knowledge should be invoked, and when. This view encourages the hypothesis that the neural representation of speech must include FPD, hence that speech is partially represented as exemplars. The debate between the storage of individual tokens versus the derivation of an abstract representation based on a set of tokens is familiar in psychology e.g. [6], and there is lively interest in episodic memory: storage of individual memory traces. However, it can be argued that, to be accessed, stored exemplars must be classified, which requires abstraction. Much psychological and phonetic evidence suggests abstract linguistic categories influence perception in many circumstances. We seek to explore the potential of models that involve both exemplar and abstract systems.

4. HUMAN SPEECH PROCESSING AND AUTOMATIC SPEECH RECOGNITION

Few of these facts are included in standard computational modelling of HSP and ASR, and despite renewed interest and promise in exemplar-based systems within the ASR community, [3, 14], such non-mainstream new systems need better psychological and linguistic frameworks to outperform state-of-the-art HMMs. To use FPD effectively in traditional computational models requires radical changes in modelling techniques, including: real speech as input, rather than 'clean' abstract categories like features and phonemes; neurophysiological and psychophysical plausibility; a radically different linguistic model from the standard; and informed searches covering long as well as short time spans. Most such needs are also relevant to exemplar-based models. The needed changes involve significant technical challenges.

For example, describing and modelling the temporal distributions of linguistic categories challenges all fields. Traditionally, speech 'segments' are seen as having short temporal domains, and prosodic categories long ones. But this simple short-long distinction is no longer workable. Properties of some segments stretch over many syllables, and intonational and rhythmic variables, carried by segments, have attributes that unfold quickly and are tied to specific places in a segment, defined by syllable type. Even in careful speech, to conducting successful conversations demands tracking non-adjacent properties of the signal. Casual speech, in which traditional segments coalesce in complex but lawful ways, can often only be understood in long contexts [4]. To derive short and long units from real speech input is an unsolved technical challenge that requires computer scientists and engineers to work closely with linguists and phoneticians.

The current situation, then, is that FPD is established as relevant to speech understanding, but poorly documented even in English, and especially in other languages. At the same time, FPD is exciting interest in speech technology and HSP modelling, as these disciplines seek solutions

to the theoretical and practical impasse produced by imposing standard linguistic theoretical constructs (e.g. early abstraction, non-redundancy) on psychological or engineering models and applications. But progress in FPD research is slow because we lack automated methods. Thus, phoneticians need computational and statistical knowledge/skills that computational modellers possess, while HSP and ASR modellers await information about FPD.

In this situation, FPD is rapidly acquiring 'cult status'. We know that it can influence perception. But we do not know that it normally influences perception, and we desperately need to establish whether/when it normally does, before it is widely taken up in a non-rigorous way as the solution to all theoretical and practical problems in speech processing. Computer science and engineering, arguably most in need of new approaches, can provide the much-needed rigour, cf. [19]. The main obstacle to progress in both fields is ignorance due to inter- and intra-disciplinary divisions. S2S aims to bridge these gaps.

5. CONCLUSION

In sum, S2S aims to nurture a new breed of researcher with multidisciplinary knowledge and skills to tackle the new theoretical framework offered by FPD. S2S will elaborate the role of FPD by building computational tools to discover systematically distinctive patterns in spoken language, to test their perceptual salience and functional relevance in everyday speech conditions and in demanding speech recognition and synthesis applications, and to construct new computational and psychological models which use long time-span speech structures. Its multilingual focus permits investigation of general patterns of FPD across structurally different languages.

The papers in this session represent part of the work of S2S, and provide a partial picture of where we are starting from: in corpus analysis [20], evaluation of episodic methods in ASR [15], prosodic-segmental interactions [18] and FL perception of non-native words in noise [12], In four years' time, we hope to be able to provide a very different picture.

6. REFERENCES

[1] Bradlow, A.R., Nygaard, L.C., Pisoni, D.B. 1999. Effects of talker, rate and amplitude variation on recognition memory for spoken words. *Perc. & Psych.* 61, 206-219.

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

[3] de Wachter, M., Demuynck, K., Van Compernolle, D., Wambacq, P. 2003. Data driven example based continuous speech recognition. *Proc. European Conf. Sp. Comm. & Technol.*, 1133-1136.

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

[5] Garcia Lecumberri, M.L., Cooke, M.P. 2006. Effect of masker type on native and non-native consonant perception in noise. *J. Acoust. Soc. Am.* 119, 2445-2454.

[6] Goldinger, S.D. 1998. Echoes of echoes? An episodic theory of lexical access. *Psych. Rev.* 105, 251-279.

[7] Hawkins, S., Slater, A. 1994. Spread of CV and V-to-V coarticulation in British English: Implications for the intelligibility of synthetic speech. *ICSLP94* Tokyo,57-60.

[8] Hawkins, S., Smith, R.H. 2001. Polysp: A polysystemic, phonetically-rich approach to speech understanding. *Ital. J. Phon.-Rivista di Ling.* 13, 99-188. http://kiri .ling.cam.ac.uk/sarah/TIPS/hawkins-smith-01.pdf

[9] Hawkins, S., Nguyen, N. 2004. Influence of syllable-coda voicing on the acoustic properties of syllable-onset /l/ in English. *J. Phonetics* 32, 199-231.

[10] Heid, S., Hawkins, S. 1999. Synthesizing systematic variation at boundaries between vowels and obstruents. *Proc.14th ICPhS* San Francisco, 511-514.

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

[12] 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.

[13] Local, J.K. 2003. Variable domains and variable relevance: Interpreting phonetic exponents. *J. Phonetics* 31, 321-339.

[14] Maier, V., Moore, R.K. 2005. An investigation into a simulation of episodic memory for automatic speech recognition. *Interspeech 2005*, 5-9.

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

[16] Nygaard, L.C., Pisoni, D.B. 1998. Talker-specific learning in speech perception. *Perc. & Psychophys.* 60, 355-376.

[17] Ogden, R.A. 1999. A declarative account of strong and weak auxiliaries in English. *Phonol.* 16, 55-92.

[18] Post, B., d'Imperio, M., Gussenhoven, C. 2007. Fine phonetic detail and intonational meaning. *Proc. 16th ICPhS* Saarbrücken.

[19] Scharenborg, O., Norris, D.G., ten Bosch, L., McQueen, J.M. 2005. How should a speech recognizer work? *Cog. Sci.* 29, 867-918.

[20] Volín, J., Studentovský, D. 2007. Normalization of Czech vowels from continuous read texts. *Proc. 16th ICPhS* Saarbrücken.

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

[22] West, P. 1999b. Perception of distributed coarticulatory properties of English /l/ and /r/. *J. Phonetics* 27, 405-426.