

Louis ten Bosch, Annika Hämmäläinen, Bert Cranen, Lou Boves
Centre for Language and Speech Technology (CLST), Radboud University Nijmegen
P.O. Box 9103, 6500 HD Nijmegen, The Netherlands
email: {l.tenbosch, a.hamalainen, b.cranen, boves}@let.ru.nl

FROM SPEECH SOUNDS TO SYMBOLIC REPRESENTATION – ABOUT NEW APPROACHES TO IMPOSE STRUCTURE ON SPEECH DATA

During the last decades, phonetic and phonological knowledge about the structure of speech has provided a basis for the development of computational models for speech recognition, including automatic speech recognition (ASR). Since speech sounds are highly variable, the robustness of the mapping from a speech signal to its discrete symbolic representation is one of the most difficult problems in speech science. The mainstream approach to ‘sound-to-symbol’ mapping is based on the use of a small set of phonetic-phonologically motivated ‘speech units’, in combination with a statistical description of these units (obtained on a large speech corpus). In this approach, speech is considered a process that can adequately be represented by a sequence of such speech units (‘beads-on-a-string’ paradigm, [1]). The performance of computational models of speech recognition has shown that this ‘beads-on-a-string’ paradigm works reasonably well for utterances that do not deviate much from the patterns included in the training corpus ([2]).

However, it is becoming clear that this conventional data-driven approach has serious limitations. Despite the use of ever-larger speech corpora, the performance of computational models of speech recognition faces a ceiling effect and falls short of human performance by an order of magnitude, particularly due to its poor capability to cope with unseen test conditions. Recently, several researchers have suggested exploring radically new approaches to address the sound-to-symbol representation. A common factor in all these new approaches is the use of sophisticated models to better impose knowledge-based structure on raw speech data. The issue of using phonological and linguistic structure is central in several lines of current research: on the role of fine phonetic details in lexical decoding ([3]), on the relation between (symbolic) context and pronunciation variation ([4]), and on the design of computational models for human speech processing ([5]). In all these research directions, the combination of statistical data-driven techniques with phonetic-phonological structure is crucial for further improvements.

In the final paper, we describe research in this new area, based on computational models for articulatory feature representation of speech. By using these features, we obtain a rich redundant representation that is particularly useful to describe possibly asynchronously events, as a first step to go beyond the beads-on-a-string constraint. Since the articulatory-to-acoustics mapping is highly non-linear [6], we hypothesize that properties of spontaneous speech such as reduction and assimilation are more parsimoniously described in terms of articulatory freedom and constraints than by acoustic-phonetic variation. To investigate this, we study reduction phenomena in spontaneous speech (especially vowel reduction and deletions, all described in terms of manual labeling of segments) in terms of these articulatory features. Finally, we report on experiments in which the bottom-up feature information is interpreted by imposing phonological structure via so-called graphical models ([7]).

References

- [1] M. Ostendorf (1999). "Moving beyond the 'beads-on-a-string' model of speech," in *Proc. IEEE ASRU-99*, Keystone, Colorado, USA. Dec 12-15.
- [2] Moore, R.K., & Cutler, A. (2001). Constraints on theories of human vs. machine recognition of speech. In R. Smits, J. Kingston, T. M. Nearey & R. Zondervan (Eds.), *Proceedings of the workshop on speech recognition as pattern classification* (pp. 145-150). Nijmegen, MPI for Psycholinguistics.
- [3] Hawkins, S. (2003). Roles and representations of systematic fine phonetic detail in speech understanding, *Journal of Phonetics*, 31, 373-405.
- [4] A. Härmäläinen, L. Boves, L., and J. de Veth (2005). "Syllable-Length Acoustic Units in Large-Vocabulary Continuous Speech Recognition," in *Proc. SPECOM-2005*, Patras, Greece, Oct 17-19.
- [5] Scharenborg, O., Norris, D., ten Bosch, L., McQueen, J.M. (2005). How should a speech recognizer work? *Cognitive Science*, 29(6).
- [6] Badin, P., Perrier, P., Boë, L.-J., Abry, C. (1990). Vocalic nomograms: Acoustic and articulatory considerations upon formant convergences. *J. Acoust. Soc. Am.* 87, pp. 1290-1300.
- [7] Bilmes, J. (2003). Buried Markov Models: A Graphical-Modeling approach to Automatic Speech Recognition, *Computer, Speech and Language*, 17.

Session: Variation, phonetic detail and phonological modeling

Presentation preference: Poster or oral.