

ISCA ITRW
Speech Analysis and Processing for Knowledge Discovery
Aalborg, Denmark, 4-6 June 2008.



Phonetic perspectives on modelling information in the speech signal

Sarah Hawkins, Department of Linguistics, University of Cambridge, UK

This talk reassesses conventional assumptions about the informativeness of the acoustic speech signal, and shows how recent research on systematic variability in the acoustic signal is consistent with a processing model that seems biologically plausible as well as compatible with recent advances in modelling embodied visual perception and action (robotics). I will review some common assumptions about the information available from the speech signal, including the strengths and limitations of phonological features and phonemes as units for encoding information in the signal. Features and phonemes are widely regarded as essential units of human speech perception and production, and, by extension, as central to machine speech recognition and synthesis systems. Yet they are abstract theoretical constructs developed within particular schools of theoretical linguistics for quite different purposes—describing phonological systems rather than speech communication. They are not intended to describe cognitive attributes of communication, nor linguistic meaning beyond contrasts in lexical form. Not surprisingly, then, models of speech perception which map cues derived from the speech signal solely onto such phonological units can only partially account for listeners' abilities. A comprehensive model of speech communication needs a broader theoretical approach, more clearly related to cognition and the wider communicative functions of speech. Present knowledge makes this goal challenging to implement, but we now know enough to begin to discuss it.

I will give examples of variation in phonetic detail which systematically signal non-phonemic linguistic information such as the grammatical or morphological status of a stretch of sound. Other examples indicate the discourse function of the utterance. Some of these systematic differences in phonetic detail are localised in the signal, while others stretch over several syllables. Both types can make speech easier to understand, presumably by increasing the signal's perceptual coherence and/or by directly indicating linguistic structure and pragmatic functions as well as phonological form.

These data encourage the development of models that focus attention away from formal phonological contrasts in citation-form words, and towards properties of connected speech that communicate other types of meaning. A currently useful approach uses principles of declarative phonology, in particular Firthian Prosodic Analysis, which places more emphasis than other models on the communicative function of an utterance, de-emphasizes the need to identify phonemes, and uses formalisms that force us to recognise that every perceptual decision is context- and task-dependent. Prosodic and grammatical properties of utterances are formally linked in twin hierarchical structures. No one descriptive unit, or hierarchical level, is more important than any other; and properties of the physical signal can inform any unit(s) in the grammatical and prosodic hierarchies in parallel.

The model advocated simultaneously accommodates detail and abstraction. It assumes parallel streams of knowledge that drive interpretation of information in the incoming signal to maximise processing efficiency. The linguistic model alone is not sufficient, but its attributes are compatible with advances in robotics which use function-oriented, body-centred, knowledge-driven systems to do visual tasks. The combined approach may be fruitful. The approach rejects the assumption that speech processing proceeds in strictly serial order, e.g. from the lowest phonological units to higher ones. Instead, sound can be mapped onto any level of linguistic structure, in parallel or in sequence, using context-dependent probabilistic processes. The important point is that there is no predetermined or rigid sequence: the process of speech perception is governed by the properties of the particular signal in conjunction with the listener's construal of the particular situation.

While this model may not prove to be the optimal one, adopting it encourages re-evaluation of the information available from the speech signal. In particular, it allows us to focus attention onto the meaning and communicative function of speech, rather than just on phonological identity. For speech technologists, this refocusing may contribute to improved unit selection and recognition algorithms.