Acoustic Modelling and Multichannel Articulatory Databases

The task of estimating speech production parameters from the acoustic waveform is known as the acoustic-to-articulatory inversion problem. Recently, the inversion problem has gained some significance in the field of automatic speech recognition (ASR) due to a growing interest in the use of articulatory parameters, either as a supplement to or substitute for spectrally based input parameters.

Conventional ASR schemes are loosely based on acoustic theories of speech which contend that a given utterance can be constructed from a basic set of atomic units called *phones*. These phones are designed to be maximally distinct in the acoustic domain and are therefore distinguished based on their acoustic properties. Thus in conventional ASR, only two levels of representation are assumed: an acoustic level and a phonetic level.

However, the observation that much of the variation which makes ASR difficult is related to articulatory phenomena (coarticulation being one of them) has lead towards an integration of speech production models and speech recognition systems [1]. The aim of this kind of research is to replace the phonetic sequence as the principal unit of acoustic modelling with an overlapping articulatory representation, assuming that physiological gestures completely specify the acoustic speech signal.

In recent studies (cf. [3], [4]), multichannel databases such as MOCHA and EUR-ACCOR have been used for investigations in the performance equivalence of articulatory and acoustic signals. The experiments consist of training a system based on Hidden-Markov-Models (HMMs) using a typical acoustic input feature vector and comparing the recognition score with the results from an articulatory input.

My contribution to the discussion will begin with a survey of multichannel articulatory databases for application in ASR. In contrast to the acoustic/articulatory feature vectors in [3] and [4] which employ conventional parameter sharing on the phonetic level, the focus of the presentation will be on subphonetic units. Subphonetic modelling has been developed as part of the SPHINX-II system and testing focused on acoustic feature vectors only [2]. Here, an extension towards articulatory feature vectors is presented. In subphonetic modelling, each of the states in phonetic HMMs is considered to constitute one such subphonetic unit. Consequently, Markov states are extracted and examined carefully for each phone model to group parameters only at the subphonetic level; similar HMMs share a representative of the output distributions called a *senone*. The set of HMMs that share output distributions through senones is referred to as shared-distribution models (SDMs).

For acoustic models, this leads to more accurate modelling [2] as it avoids over-generalisation in the sense that pooling all the vectors of a phone into one acoustic model may result in inaccurate modelling for any particular unit. Research is currently underway to derive articulatory-driven senonic baseforms using the EUR-ACCOR multichannel database.

Moritz Neugebauer

[1] L. Deng, G. Ramsay, and D. Sun. Production models as a structural basis for automatic speech recognition. In *Proceedings of the Fourth European Speech Production Workshop*, Autrans, 1996.

[2] M. Hwang. *Subphonetic Acoustic Modeling for speaker-independent Continuous Speech Recognition.* PhD thesis, Carnegie Mellon University, 1993.

[3] A. A. Wrench. A new resource for speech production modelling in speech technology. In *Proceedings of Institute of Acoustics Workshop on Innovation in Speech Processing*, Stratford-upon-Avon, 2001.

[4] A. A. Wrench and W. J. Hardcastle. A multichannel articulatory speech database and its application for automatic speech recognition. In *Proceedings of 5th Seminar on Speech Production*, Kloster Seeon, 2000.