New Aurora Activity for Standardization of a Front-End Extension for Tonal Language Recognition and Speech Reconstruction

June 2001

The European Telecommunication Standards Institute (ETSI) STQ-Aurora group [1] has created a new work item to address the standardization of an extended front-end for Distributed Speech Recognition (DSR) of tonal languages as well as speech reconstruction for DSR systems.

In February 2000, the ETSI STQ-Aurora group published the first DSR standard with a Mel-Cepstrum front-end [2]. At present, the group is in the advanced stages of standardizing a new noise robust front-end. An overview of DSR and the work of the Aurora group can be found in [3]. A block diagram of the DSR components is presented in Figure 1.

The standard Mel-Cepstrum front-end, as well as the future noise-robust front-end, defines the extraction of spectral features on the mobile terminal. These features are compressed and transmitted to the server for recognition back-end processing. The front-end standardization process includes recognition tests performed in several European languages, as well as American English. It is well known, however, that for some Asian languages (such as Mandarin, Cantonese and Thai) recognition can be improved when tonal information is introduced in addition to spectral information. To promote the use of the ETSI DSR standards in Asia, the Aurora group has decided to lead efforts towards a standard for extracting and compressing tonal information for DSR systems. A typical deployment of DSR systems for processing tonal languages will use the standard Mel-Cepstrum or noise-robust front-end features, with the addition of standard tonal features.

Some applications may require reconstruction of the speech waveform from the DSR features (on the server side). Examples of these applications include:

- Interactive Voice Response (IVR) services based on the DSR of “sensitive” information, such as banking and brokerage transactions. DSR features may be stored for future human verification purposes or to satisfy legal requirements.
- Human verification of utterances in a speech database collected from a deployed DSR system. This database can then be used to retrain and tune models in order to improve system performance.
- Applications where machine and human recognition are mixed (e.g., human assisted dictation).

The Aurora group is also working towards standardizing the extraction and compression of additional features required for speech reconstruction. These features can optionally be transmitted from the mobile terminal to the server along with the Mel-Cepstrum or noise-robust spectral features. Since tonality is required for both speech reconstruction and for tonal language recognition, these tasks have been assigned to the same working team within Aurora.

For all the applications mentioned above, the need for intelligibility far outweighs the need for better voice quality of the reconstructed speech. Therefore, the assessment of speech reconstruction solutions will be based primarily on testing their intelligibility. It is expected that a human listening to the reconstructed speech will be able to recognize the spoken material much better than an automatic recognizer using the DSR features.

Figure 2 presents a block diagram describing the standardization approach adopted by Aurora. The fundamental-frequency ("pitch", F0) will be extracted on the terminal from the recorded voice at a fixed update rate of 10 – 20 msec. The extraction will not introduce any additional delay (look-ahead) on top of the delay already inherent in the computation of the spectral features. Furthermore, only minimal constraints will be applied in the fundamental-frequency search to allow further propriety server side processing of the "raw" fundamental-frequency. The fundamental-frequency will be compressed and transmitted during the voiced portion of the recorded speech. Voiced/unvoiced classification (V/UV) will also be transmitted.

Other features may need to be extracted on the mobile terminal and transmitted in order to:

- Enable a variety of propriety pitch tracking methods for tonal language recognition. Pitch tracking is required in order to eliminate gross pitch errors (such as double or half pitch instances) and voicing classification errors. A pitch tracking algorithm for recognition purposes will not be standardized.
- Enable pitch tracking for the purpose of speech reconstruction.
- Improve the intelligibility and quality of the reconstructed speech.

Examples of "other features" are voicing degree, sub-frame energies, etc. It is expected that no more than 1 kb/s will be required for the transmission of the fundamental-frequency, V/UV decision, and any other features. This is on top of the bits already allocated for transmission of the spectral features (e.g., 4.8 kb/s for the Mel-Cepstrum features).

Extracting accurate fundamental-frequency and other features on the terminal, under the restriction of limited processing power and memory, is a definite challenge. Moreover, accuracy must be maintained even for noisy speech recoding, which is typical of mobile applications.

For more information about this and other Aurora activities, please contact the chairman of Aurora, Mr. David Pearce (Motorola UK), e-mail: bdp003@email.mot.com

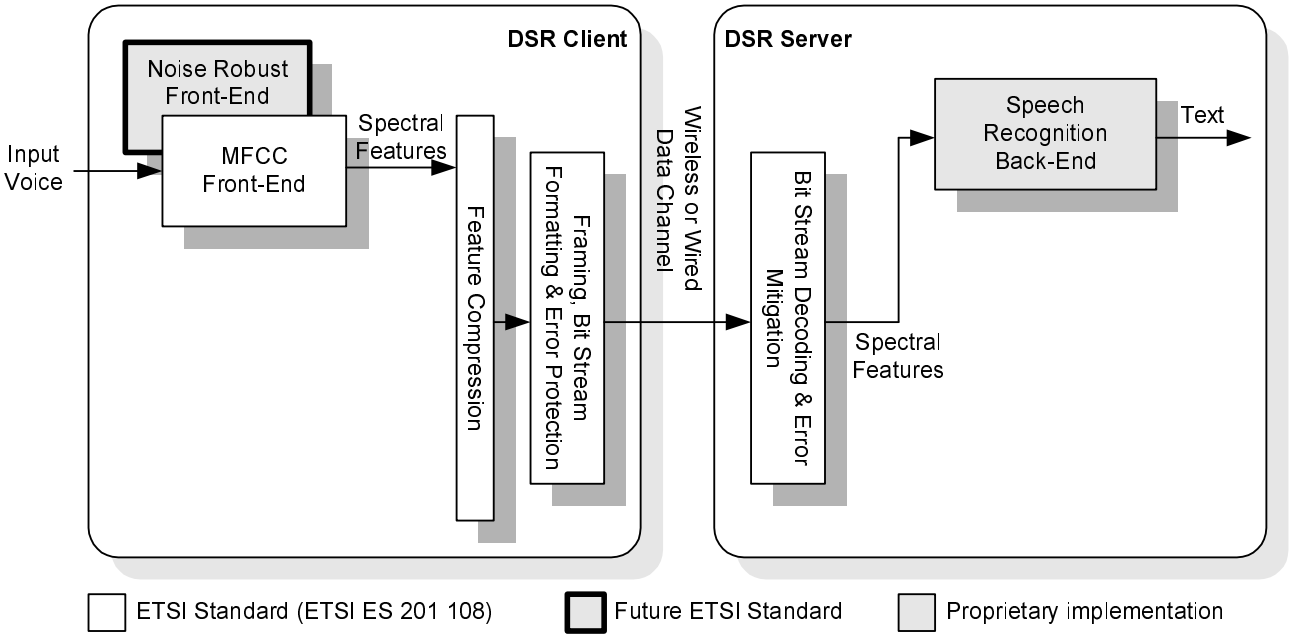

Figure 1: Components of the ETSI Aurora Mel-Cepstrum and future noise robust front-end standards for DSR applications

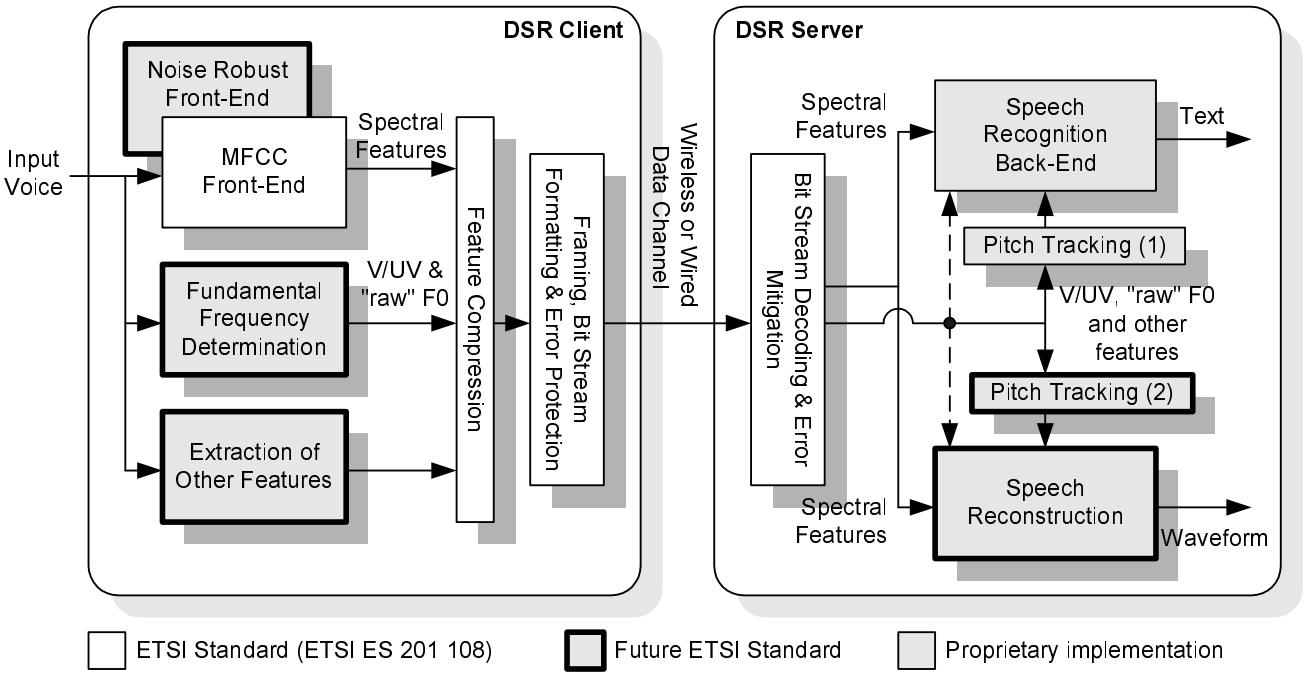

Figure 2: Components of future ETSI Aurora standards enabling DSR of tonal languages and speech reconstruction

References:

[1] http://www.etsi.org/aurora/

[2] ETSI ES 201 108 version 1.1.2, “Speech Processing, Transmission and Quality aspects (STQ); Distributed speech recognition; Front-end feature extraction algorithm; Compression algorithms”, April 2000.

[3] D. Pearce, “Enabling new speech driven services for mobile devices: An overview of the ETSI standards activities for distributed speech recognition front-ends”, in proc. American Voice I/O Society (AVIOS), May 22-24, 2000, San-Jose CA, USA.