



A REAL-TIME AUDITORY 'SPECTROGRAPH' FOR SPEECH RESEARCH

Frank Gooding* and Ian Shaw^o

ABSTRACT

A DSP-based auditory transform allowing for real-time presentation of an "auditory spectrogram" of any audio signal is described. Various parameters can be extracted from this representation and simultaneously displayed. Such a tool allows a variety of applications for research in speech perception and ASR.

INTRODUCTION

A model of peripheral auditory processing is implemented on a Texas Instruments TMS 320C25 Digital Signal Processor mounted on an IBM PC-AT compatible microcomputer. The TMS 320C25, running at 40 MHz, performs both the spectral analysis of a sampled signal (fast Hartley [ref 1] and Walsh transforms are currently being tested), and the auditory transform at a speed that allows for near real-time presentation of an "auditory spectrogram" of any audio signal. The complete auditory model described in more detail in elsewhere (ref 2, 3) is based on that of Moore and Glasberg (ref 4, 5), but with the addition of conversion of the ordinate to loudness level (phon conversion) and loudness scaling (sone conversion).

MODEL AND IMPLEMENTATION

Computation of the auditory transform takes place in four stages. First, a 512 point log power spectrum of the sampled signal is computed every 10 msec. Next, The abscissa in Hz is converted to the perceptually relevant ERB (equivalent rectangular bandwidth) scale, reflecting the frequency resolving properties of the peripheral auditory system (ref 4) and quantized to .1 ERB between 3 and 30 ERB. The result is a dB versus ERB representation. In the current, streamlined version of the model, the ordinate of this intermediate stage is at this point converted to loudness level in phons, reflecting the frequency sensitivity of perceived loudness. This is performed simply by table look up, appropriate values for every .1 ERB having been interpolated from the standard tables (ref 6). The resulting representation is in the fourth stage convolved with the auditory filter function, a rounded exponential (ref 7).

The demand for speed of throughput for the present application has meant that some features of the complete

*Dept. of Linguistics, Univ. Coll. N. Wales, Bangor, LL57 2UW
^oSchool of Electronic Engineering Sci., UCNW.

model, seen as refinements which substantially increase the computational load, have been provisionally omitted for the present. One noteworthy feature to fall victim to this streamlining is the level sensitivity of the auditory filter shape. In its basic form, the filter shape is symmetrical about the center, but at levels above 51 dB SPL, the filter becomes asymmetric, the slope of the HF skirt decreasing with increasing intensity (ref 5). Since the level sensitivity is expressed as a function of input intensity in dB SPL, conversion to phons must follow convolution with the (variable) auditory filter. This is not required in the streamlined model with constant filter shape. A further stage of the complete model, that of loudness scaling (some conversion), following the formula of Zwicker and Scharf (ref 8), is also omitted for computational savings, leaving the ordinate scale in dB (which is at any rate a good perceptual approximation). Trials of recognition accuracy with and without the omitted features are planned.

The main time savings accruing from the use of the DSP over conventional microprocessor implementations come in the calculation of the spectral transform and the auditory filter convolution, both extremely computationally demanding for conventional microprocessors. In addition, use of a separate DSP board allows for parallel processing, since the computer's CPU (the Intel 80286 in this case) is freed to perform additional analysis of the transformed signal while the DSP is performing the transform operations.

APPLICATIONS

A variety of display modes are available to aid researchers in examining and extracting information from the auditory representations. These include spectrogram, waterfall, or section. "Spectrogram" mode will display pitch versus time, with loudness shown by gray level (currently mapped onto a scale of 16 levels). Individual spectral sections (as on a traditional spectrograph) can be chosen for detailed display by a movable cursor. Various parameters can be extracted from such representations and simultaneously displayed. Since it is implemented in software, alterations of model characteristics are easily made.

The system described here should serve as a valuable and flexible tool for speech research. At relatively small expense, it allows speech scientist to investigate the auditory properties of speech sounds with an ease and rapidity previously unobtainable even on minicomputers. In addition, the DSP-based auditory transform presented here also aims significantly to improve the accuracy and efficiency of the acoustic-to-phonetic front end of an ASR system. Improvement in this area has been accorded a high priority by ASR researchers in recent years, as it is seen as essential to enhancing overall system performance. The basic rationale for an auditory conversion in an ASR front end is that it allows one to focus on the perceptually salient

features of the speech signal for parameterization, thereby increasing efficiency and accuracy. Preliminary work towards extracting robust auditorily-based parameters including sonority, sharpness, and vowel features is in progress. Use of such auditory displays are also expected to contribute to solving the problem of speaker normalization.

References:

1. R. Bracewell Proc. IEEE 22.8, 1010 (1984)
2. F. Gooding JASA 80:Suppl. 1, S126 (1986)
3. F. Gooding, Bangor Res. Papers in Ling. 2, 27 (1987)
4. B. Moore and B. Glasberg JASA 74.3, 750-753 (1983)
5. B. Moore and B. Glasberg, in Moore, B.C.J., ed., Frequency Selectivity in Hearing, London, Academic Press Chapt. 5 (1986)
6. D. Robinson, R. Dadson Brit. J. App. Physics 7, 166 (1956)
7. R. Patterson et al JASA 72, 1788-1803 (1982)
8. E. Zwicker, and B. Scharf, Psych. Rev. 73, 2 (1965)