# SIGNAL PROCESSING IN SPEECH AND HEARING TECHNOLOGY

**Sean A. Fulop**
*Dept. of Linguistics, California State University, Fresno*
*Fresno, California 93740*

**Kelly Fitz**
*Signal Processing Group, Starkey Laboratories*
*Eden Prairie, Minnesota 55344*

**Douglas O'Shaughnessy**
*National Institute of Scientific Research, University of Quebec*
*Quebec City, Laval, Montreal, H5A 1K6, Canada*

Speech science and technology would scarcely exist today without acoustic signal processing. The same can be said of hearing assistance technology, including hearing aids and cochlear implants. This article will highlight key contributions made by signal processing techniques in the disparate realms of speech analysis, speech recognition, and hearing aids. We can certainly not exhaustively discuss the applications of signal processing in these areas, much less other related fields that are left out entirely, but we hope to provide at the very least a sampling of the wide range of processing techniques that are brought to bear on the various problems in these subfields.

*"Multiband compression is the core of modern digital hearing aid signal processing, and is the primary tool for restoring audibility and comfort to patients with hearing loss."*

While speech itself is an analog signal (or time sequence) of air pressure variations resulting from puffs of air leaving one's lungs, modulated by the vibrations of one's vocal cords and filtered by one's vocal tract, such a vocal signal is normally digitized in most modern applications, including normal telephone lines. The analog-to-digital (A/D) conversion is needed for computer processing, as the analog speech signal (continuous in both time and amplitude), while suitable for one's ears, is most efficiently handled as a sequence of digital bits. A/D conversion has two parameters: samples/second and bits/sample. The former is specified by the Nyquist rate, twice the highest audio frequency to be preserved in the speech signal, assuming some analog filter suppresses the weaker energy at relatively high frequencies in speech (e.g., above 3.3 kHz in telephone applications, using 8000 samples/s). Like most audible sounds, speech is dominated by energy in the lowest few kHz, but pertinent energy exists to at least 20 kHz, which is why high-quality recordings, such as CDs, sample at rates up to 44.1 kHz. However, speech can be reasonably intelligible even when low-pass filtered to 4 kHz, as the telephone amply demonstrates. Typical speech applications use 16-bit A/D accuracy, although basic logarithmic coding in the telephone shows that 8-bit accuracy can be adequate in many applications, which include automatic speech recognition, where the objective is a mapping into text, rather than a high-quality audio signal to listen to or analyze in depth.

## Speech spectrum analysis

In phonetics and speech science a commonly pursued aim is to analyze the spectrum of speech as completely as possible, to obtain information about the speech articulation and the specific auditory attributes which characterize speech sounds or "phonemes" (consonants and vowels). Spectrum analysis can further our understanding of the variety of sounds in language (a linguistic pursuit), and can also further our understanding of the fundamental nature of normal and disordered speech.[1]

### Fourier spectrum and spectrogram

The simplest form of spectrum analysis is the Fourier power spectrum with which readers are probably familiar. Although this is not a function of time, and therefore involves an assumption of stationarity across the signal frame analyzed, it is still of some utility for speech analysis. The Fourier spectrum is particularly useful for examining the spectral characteristics of speech sounds whose steady state is very important, such as fricative consonants (e.g., 's' and 'sh'). Because such sounds are chiefly noisy, it can be useful to apply statistical techniques such as ensemble averaging to examine their spectra.

The natural extension of the Fourier transform to the time-frequency plane is provided by the *short-time Fourier transform* (STFT), which, in digital form, is essentially a time-frequency grid of complex numbers. Each frequency column of this matrix at a given time point is the discrete Fourier transform of the analysis window on the signal at that time. The log magnitude of the STFT matrix is traditionally called the *spectrogram*, and has long been a popular means of examining the spectrum of speech as it changes through time. Figure 1 shows a speech signal waveform, together with a very brief window cut from the signal during the vowel. It is this type of short window, suitably tapered (e.g., by multiplication with a Gaussian function), that can be used to create a spectrogram using power spectra of successive overlapped windows as shown in Fig. 2.

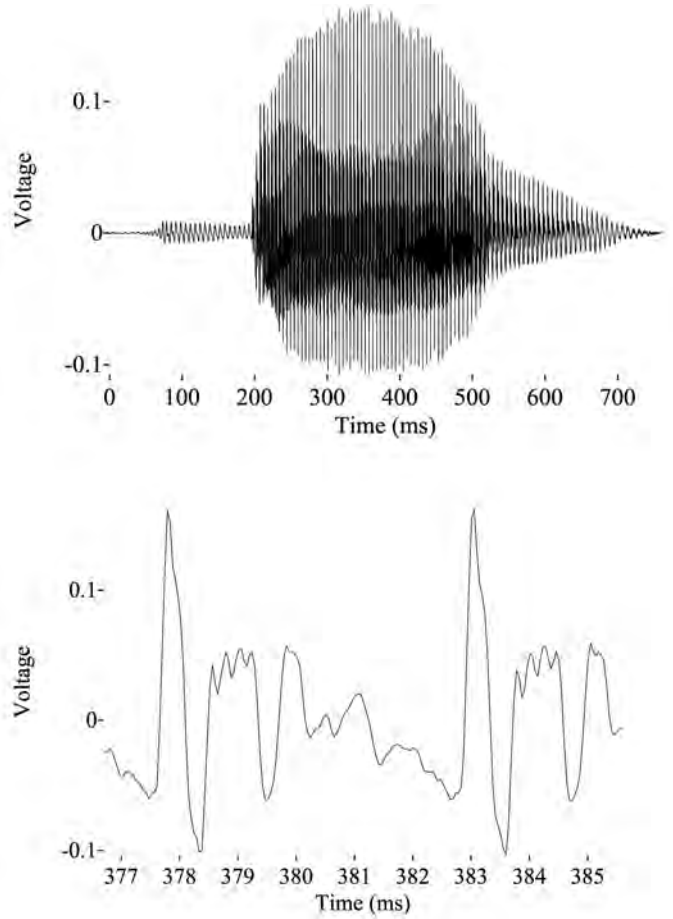


*Fig.1. Signal waveform of Somali word [bal] (top panel) and a 9 ms analysis window taken from the middle of the vowel.*

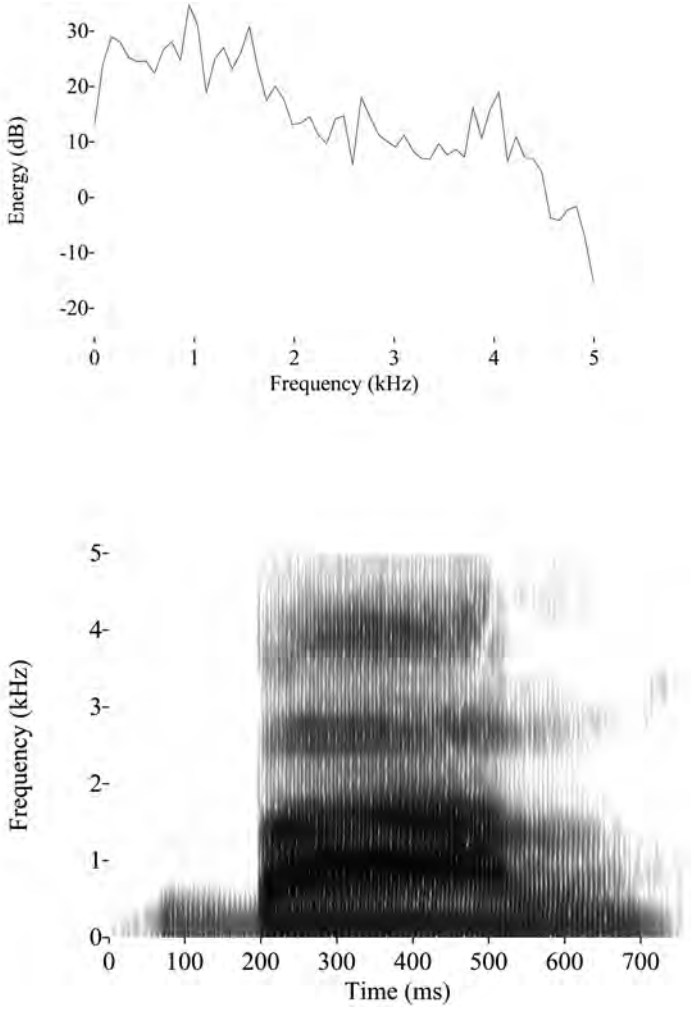


*Fig. 2. Power spectrum (upper panel) of one analysis window (v. Fig. 1) after Gaussian tapering; spectrogram of Somali word [bal] created using power spectra of successive overlapping windows. Spectrogram amplitude is shown using a grayscale linked to decibel values, with the darkest areas representing the highest spectral energy.*

## Quadratic time-frequency analysis

The spectrogram is but one example of a wide range of *time-frequency representations* which aim to show the distribution of signal energy over the time-frequency plane in the most useful way. There are so many options now that a huge literature on the subject has developed over the past thirty years or so, ever since the Wigner-Ville distribution (WVD) was given a practical algorithm in the digital setting by Claasen and Mecklenbräuker.[2] A time-frequency representation is, roughly speaking, a sort of "running spectrum analysis," a spectrum that also changes through time. A spectrogram takes this definition and uses it literally, providing a time sequence of power spectra. But the WVD and its generalizations come at this idea from a different and more subtle angle.

The WVD is most easily understood in relation to the signal autocorrelation, which may be defined in the following way for a digital signal $s(n)$ consisting of $N$ samples:

$$r(l) = \sum_{l}^{N-1} s(n)s(n-l) \qquad (1)$$

Observe that the autocorrelation is a function of $l$, known as the lag time; it is basically the result of multiplications of the signal with time-shifted copies, and so it has large peaks at those lags where the signal closely matches itself, thereby serving to detect periodicity. An important fact about the autocorrelation of a signal is that its Fourier transform yields a power spectrum directly.

Normally, the WVD is defined on a complex counterpart of a signal called its analytic associate, which we denote by $z(n)$. At the core of the definition is a "running" version of autocorrelation called the *instantaneous autocorrelation*, defined as a function of both the time and lag variables:

$$K(n,l) = \sum_{|l|<N/2} z(n+l)z^*(n-l) \qquad (2)$$

The WVD is then basically just the Fourier transform of this; it provides a kind of running spectral analysis for the same reason that the stationary spectrum is obtained by Fourier transform from the standard autocorrelation. Cohen[3] originally showed how the WVD can be further generalized to a class of related representations by convolution with an additional smoothing kernel, defining the "Cohen class" of quadratic time-frequency representations. The purpose of the smoothing kernel is to smooth out the various interference terms that pollute the WVD and clutter the image with unphysical information. One of the most useful examples of the Cohen class, the Zhao-Atlas-Marks distribution,[4] is illus-

trated in Fig. 3 as an alternative to the standard spectrogram for providing cleaner analyses of speech data.

**Reassigned spectrogram**

The STFT underlying the digital spectrogram is a grid of complex points in the time-frequency plane, each constituting a magnitude and phase angle. The spectrogram itself is displayed using only the magnitude, converted to a decibel grayscale; the phase is discarded. However, there is much useful information in these STFT complex phases, which can be harnessed to compute the instantaneous frequency and precise time instant corresponding to the spectrographic grid points. Using this information, it is possible to reassign the points in a spectrogram to new locations in the time-frequency plane. A reassigned spectrogram can then be displayed as a 3-D scatterplot showing all these points replotted in time and frequency, but using the original spectrographic magnitude. Once again it proves effective to display the magnitude dimension by means of a colormap.

The instantaneous frequencies of the signal components are determined from the time derivative of the STFT phase, while the correct time instants for each excitation are determined from the frequency derivative.[5] These calculations are faithful to the original spectrogram, and so, unfortunately, the interference terms are also reassigned and displayed. It is possible to extract yet more information from the STFT phase, by computing the higher-order mixed partial derivatives.[6] These quantities can be used to determine whether a given point in the reassigned spectrogram is closely affiliated to a signal component, an impulsive event, or is dispensable because it is probably not affiliated to a signal element in any realistic sense. A reassigned spectrogram which has points removed by employing these derivatives has been called *pruned*; examples of the results are shown in Fig. 4. There it can be seen that, because of the increased precision in locating instantaneous frequencies, it is valuable to use a reassigned spectrogram of a very brief signal segment. Formant frequencies (vocal tract resonances) of a vowel are much more easily located and measured from this kind of "magnified" analysis, because such quantities appear to vary within each cycle of the vocal cords. The detailed view of the vowel in Fig. 4 is greatly affected by the speaker's voice quality, such that the important formant frequencies which characterize the vocal tract shape now have to be separated from other resonances resulting from coupling to the trachea.

## Speech coding and automatic speech recognition

Humans have often sought to design machines to accomplish tasks to emulate their own behavior, i.e., artificial intelligence. One popular application that is increasingly

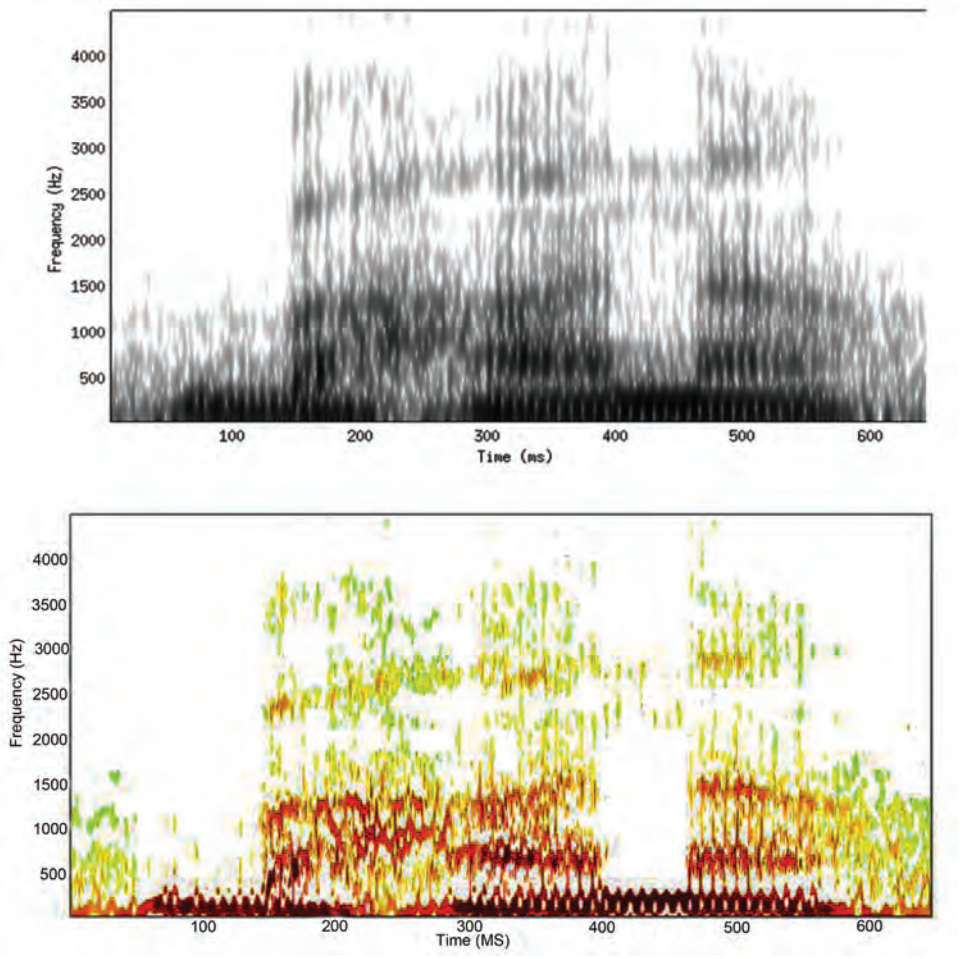


*Fig. 3. Hindi word [bhana] shown with a spectrogram (upper panel) and a Zhao-Atlas-Marks image computed with the same analysis window length of 9 ms. The colormap runs from dark red (loudest) to green (quiet) in standard order.*

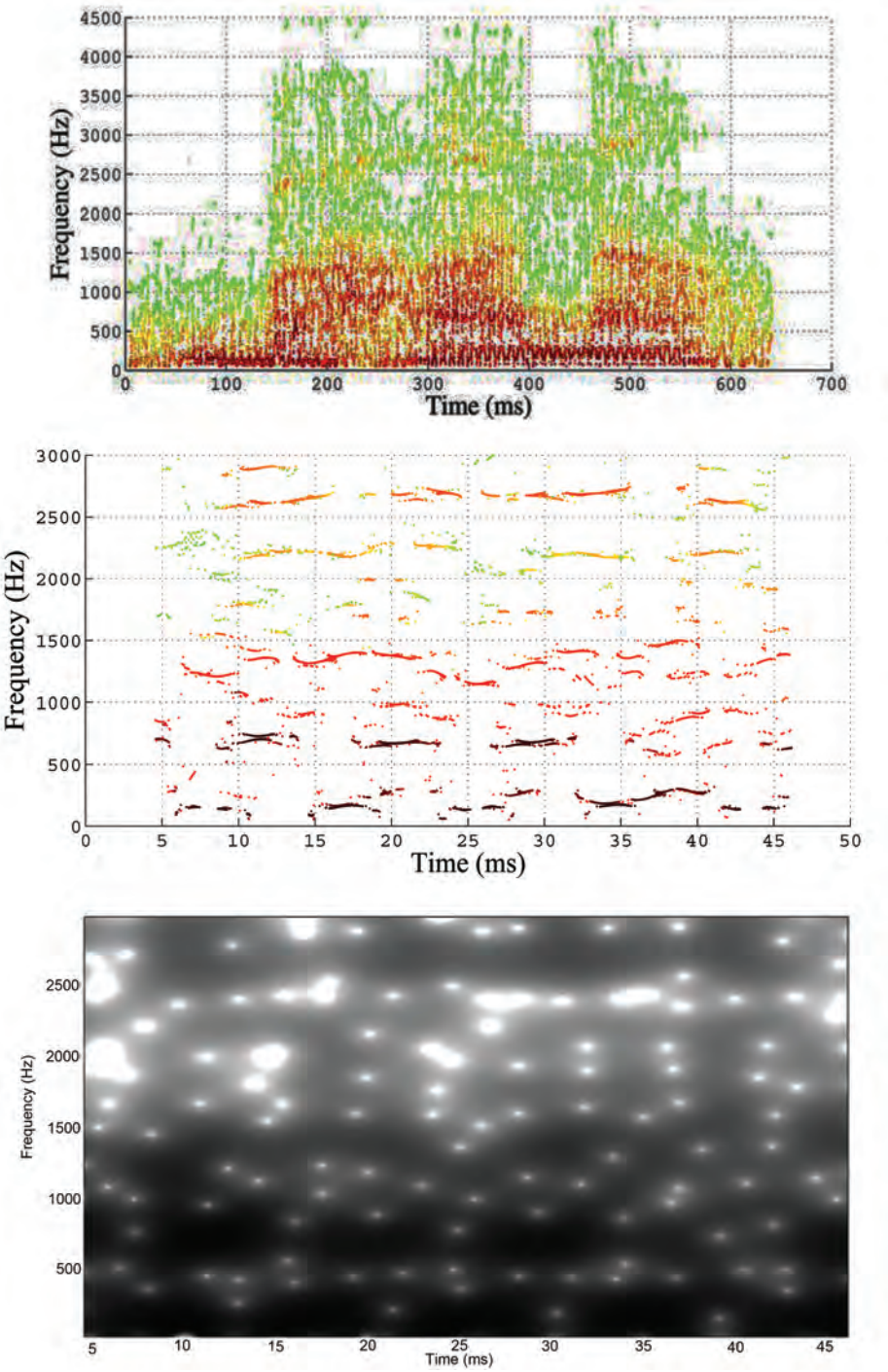


*Fig. 4. Hindi word [bhana] shown using reassigned spectrogram (upper panel) pruned to remove noise and irrelevant points. Middle panel shows a reassigned spectrogram of a brief portion of the first vowel [a], pruned to show frequency components only. Lower panel shows the corresponding conventional spectrogram with the same analysis windows (9 ms).*

found in telephone interfaces is human-to-computer voice dialog, where people can access information and effect transactions verbally without needing a human operator. This requires both automatic speech recognition (ASR), to convert one's voice into a textual message without manual assistance, and text-to-speech (TTS), to formulate the verbal responses. Both of these tasks are founded on algorithms that involve digital signal processing. We will here discuss chiefly the recognition aspect, but speech coding has often used many of the same processing techniques as speech recognition "front ends."

Speech coders, as found in modern cell phone technology, often use a linear prediction (LP) approach.[1,7] LP estimates each sample of a speech signal based on a linear combination of a small number (e.g., 10) of its immediately preceding samples. The speech signal is thus modeled statistically as an autoregressive process. Such a model also determines a digital filter representing the vocal process, from standard tenets of filter theory. The multiplier weights in the resulting filter allow simple synthesis of speech very effi-

ciently, and also allow efficient representation of the state of one's vocal tract in the form of a vector with on the order of ten parameters per frame of speech data. Ten is sufficient, as speech typically has one resonance per kHz and each resonance is specified by two complex numbers. Each frame of speech is typically 10 ms in an ASR application, which represents a compromise duration, being long enough to contain a sufficient number of samples, while avoiding averaging out dynamic movements of the vocal tract over long periods. It may be mentioned that LP analysis is frequently used to provide a concise account of the formants for phonetic analysis, which is much less labor-intensive than the spectrum analysis procedures outlined in the previous section, but may also be considerably less accurate.

ASR[8] is an example of a pattern recognition task, where one maps an auditory object (the speech signal) into a classification (text). This involves data compression, as the initial object is usually represented with an extensive bit sequence, while the output classes are far smaller in number. For speech, typical compressions are many orders of magnitude, from, say, 32 kbits (for a half-second uttered word using telephone's logarithmic pulse-code modulation) down to as little as one bit (e.g., a simple vocabulary of yes versus no), a few bits (for a digit recognition application), or perhaps 10-20 bits (to accommodate the hundreds of thousands of possible words in a given language). Ideally, any data compression would eliminate less useful information, while retaining detail pertinent to discriminate the relevant classes of words in an allowed vocabulary (e.g., assuming that a speaker utters one word at a time). Early ASR simply used the Fourier transform spectrum mentioned above, but this does little significant compression, only allowing us to discard the phase.

The cepstrum is defined as the inverse Fourier transform of the power spectrum of input speech.[9] The phase is discarded as being of little use so far in ASR, as it typically reflects details of three-dimensional air flow in the vocal tract (VT), while ASR is concerned about the shape of the VT, as the latter reflects what sounds and words one is uttering. The VT shape correlates strongly with positions of peaks in the speech spectra, e.g., one's F1 (first "formant" or resonance) varies directly with tongue height, while F2 varies with front-back tongue location and lip rounding. The logarithmic amplitude compression relates to the normal nonlinear compression that appears in much of human perception, whether touch, sound or vision. The cepstrum is a kind of "spectrum of the spectrum," and as such it factors out information about the peaks in the power spectrum while providing a set of decorrelated components, called *cepstral coefficients*, representing the signal. See the above section on reassignment for some new developments on harnessing the phase to improve precision in determining the resonance locations.

For modern ASR, the most common data representation is the mel-frequency cepstral coefficients (MFCC). For ASR purposes, the LP and standard cepstral representations have a weakness in their treatment of all frequencies as equally important, as does the Fourier transform with its fixed bandwidth, unlike, say, wavelets. The mel scale is a non-linear mapping of physical frequency to a perceptual scale, following the logarithmic arrangement of frequency "bins" along the basilar membrane in the human cochlea. *A priori*, it is not obvious that ASR needs to follow such aspects of human hearing, but empirical evidence of superior recognition accuracy with MFCC[10] has led to its acceptance in ASR.

The final step of the MFCC, the inverse Fourier transform, allows capturing the essence of the VT shape in as few as 10-16 parameters, as the cepstrum effectively converts convolution to addition. Speech is often viewed as the filtering of a glottal waveform by the frequency response of the VT; hence its spectrum is the product of that of the glottis and the VT. As the log of a product is the sum of its log components, the cepstrum conveniently consists of a linear combination of the desired VT component and the undesired glottal one. Furthermore, the former is compressed in the low end (as it represents spectral envelope details, i.e., the resonances, which vary slowly in frequency), in terms of the first 10-16 samples, while the latter is at the high end of the cepstrum, as it reflects the harmonics, which demonstrate rapid amplitude variation at multiples of the speaker's vocal cord vibration rate. As ASR seeks VT information and not glottal detail, the cepstrum is a convenient way to discard the glottal effects by simply employing the first few coefficients. Figure 5 shows a word represented with a spectrogram, and a similarly processed image which lays out the mel-frequency cepstral coefficients along the frequency axis. It is striking how

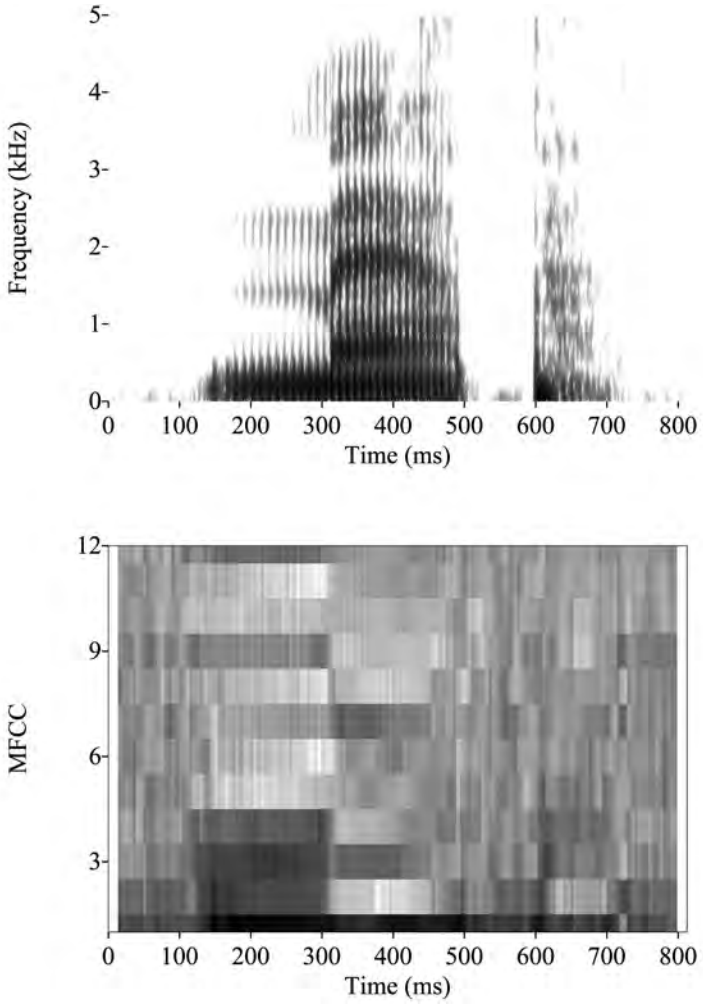


*Fig. 5. English word [map] shown in a spectrogram and a corresponding layout of the mel-frequency cepstrum coefficient matrix.*

the phonemes in the word are separated by sharp boundaries in the MFCC "spectrogram," a quality that is no doubt very helpful for the speech recognition pattern-matching which would normally use such a representation.

Other parameters have been used for ASR in the past, e.g., fundamental frequency (F0), energy, zero-crossing rate, and autocorrelation. They are still used in many other speech applications, such as coding and speaker verification. To estimate F0, one looks for cues to the periodic vibration rate of the vocal cords in the corresponding speech signal. Of course, speech is never truly periodic, but only quasi-periodic, as a speaker alters F0 for a myriad of purposes (syntactic structuring, emphasis, tones in a tone language, emotion). The major point of excitation of the VT occurs at vocal cord closure, after which the speech energy decays. So simple peak picking of the speech signal $s(n)$ is a direct way to estimate F0. However, signal processing enhances accuracy; examples include using autocorrelation, which leads to clearer peaks than found in the original $s(n)$. This owes much to the fact that phase is eliminated in the autocorrelation.

Energy, being simply the sum of the square of a sequence of speech samples over a frame, is a basic useful measure for many speech applications, such as voice activity detection. Use of the square operation, rather than some other power, is justified heuristically and by its use in Parseval's Theorem (stating that the energy in a signal is completely specified by the squared Fourier transform). Very-low-rate speech coders transmit energy, F0, and the LP multiplier coefficients every frame, at typical rates of 2.4 kbits/s, although modern cell phones also send information about phase as well, leading to higher quality at rates of 8-11 kbits/s.

The zero-crossing rate is a very simple measure of spectral prominence. Just by counting the number of times the speech signal changes algebraic sign (e.g., as the air pressure in front of the mouth goes from exceeding the ambient atmospheric pressure to being less), one gets a good estimate of what frequency dominates the energy. For example, a sine wave has two crossings per period.

All the above measures have found utility in speech applications for purposes of data compression, converting a very high bit rate signal sequence into more efficient parameter sets. In ASR and coding, further processing is possible by the use of delta parameters, i.e., calculating the difference between successive samples in time. Delta modulation coders exploit the fact that most audio signals are dominated by low frequency energy. One still needs to preserve higher frequencies, but D/A quantization noise is less when using differenced parameters. For ASR, we use the difference of successive frames to represent the velocity of vocal tract movements, placing such information (and, in some cases, a double difference, to model acceleration) into a single vector per frame of data, so as to accommodate the first-order assumption of the standard hidden Markov models which provide the typical ASR search method. Yet another example of differencing is that of Cepstral Mean Subtraction (CMS), in which one may subtract the long-term average spectra from that in each frame, before applying the data to ASR. Just as the similar Dolby processing can suppress tape hiss, CMS attempts to suppress channel and noise characteristics, while preserving the more relevant dynamic aspects of vocal tract movement for speech recognition.

## Signal processing in hearing aids

### Compression

One of the symptoms of sensorineural hearing loss is a shift in hearing thresholds that produces an overall reduction in loudness, especially for quiet sounds, and an abnormal growth of loudness that causes loud sounds to be very loud, and quiet sounds to be inaudible. The loss of dynamic range compression provided by a healthy cochlea produces a condition known as *recruitment*. Recruitment reduces the dynamic range over which hearing-impaired listeners can perceive sound, so listeners with hearing loss may find quiet sounds inaudible and loud sounds painful. Consequently, straightforward linear amplification, making all sounds louder, is not an effective treatment for most patients.

Modern digital hearing aids apply *wide dynamic range compression* to treat the abnormal growth of loudness due to recruitment. Compression, a form of automatic gain control, amplifies quiet sounds more than loud sounds, reducing the overall dynamic range of the processed sound, and allowing

quiet sounds to be made audible without making loud sounds uncomfortable. For patients suffering the reduced dynamic range that is typical of sensorineural hearing loss, compression can provide audibility and comfort over a wider dynamic range than linear amplification.[11]

In *wideband compression*, gain is computed according to the overall signal level, and applied equally across all frequencies. This technique preserves the spectral shape of the processed signal, but it has the disadvantage that the gain computation is dominated by the region of the spectrum having the greatest energy. The presence of a strong, narrowband signal in one frequency region can thereby cause weaker signals at distant frequencies to be rendered inaudible. Moreover, most patients suffer hearing loss that is non-uniform in frequency. Wideband compressors offer no means of prescribing more compression in frequency regions of greater hearing loss.

In *multiband compression*, the signal is filtered into several frequency bands, by means of a filterbank or a discrete Fourier transform, and compression is applied independently to the signal in each band. At any one time, the gain applied in a multiband compression prevents a strong, narrowband signal from triggering a gain reduction at distant frequencies, and allows compression to be prescribed differently in each band according to the patient's hearing loss. As many as 32 bands of compression may be employed in a hearing aid, depending on the manufacturer.

Compressive amplification is described by the ratio of input level (in decibels) to output level. A compression ratio of 2:1, for example, implies that a change in input level of 2 dB produces a 1 dB change in output level. Compression ratios in hearing aids rarely exceed 3:1. Typically, sounds quieter than the *compression threshold* receive linear amplification (1:1), to avoid excessive amplification of low level background noise, and in some cases, they may be attenuated (this is called dynamic range *expansion*).

Instantaneous gain changes introduce artifacts and compromise sound quality, so compression circuits are further characterized by a pair of time constants that determine how quickly the gain is reduced when a sudden increase in signal level is detected (the *attack* time constant) and how quickly the gain is restored due to a drop in signal level (the *release* time constant). Attack time constants are often short, on the order of tens of milliseconds or less, to prevent a sudden loud sound being presented with painfully high gain to the hearing aid wearer. Release time constants are typically tens to hundreds of milliseconds. Time constants are often uniform across all bands, but need not be so.

While many people with hearing loss experience loud sounds similarly to people with normal hearing, a further consequence of hearing loss for some patients is an increased sensitivity to loud sounds, called *hyperacusis*. To prevent very loud sounds from causing discomfort or saturation, hearing aids may employ a further stage of heavy *output limiting* compression to keep the output signal within "safe" limits.

Multiband compression is the core of modern digital hearing aid signal processing, and is the primary tool for restoring audibility and comfort to patients with hearing loss. In addition, many other signal processing algorithms are employed to increase patient comfort, to improve sound quality, to provide microphone directionality, or to treat special conditions of hearing loss. For example, dedicated signal processing is applied to the problem of restoring audibility of high-frequency speech cues to patients with severe to profound high frequency hearing loss.

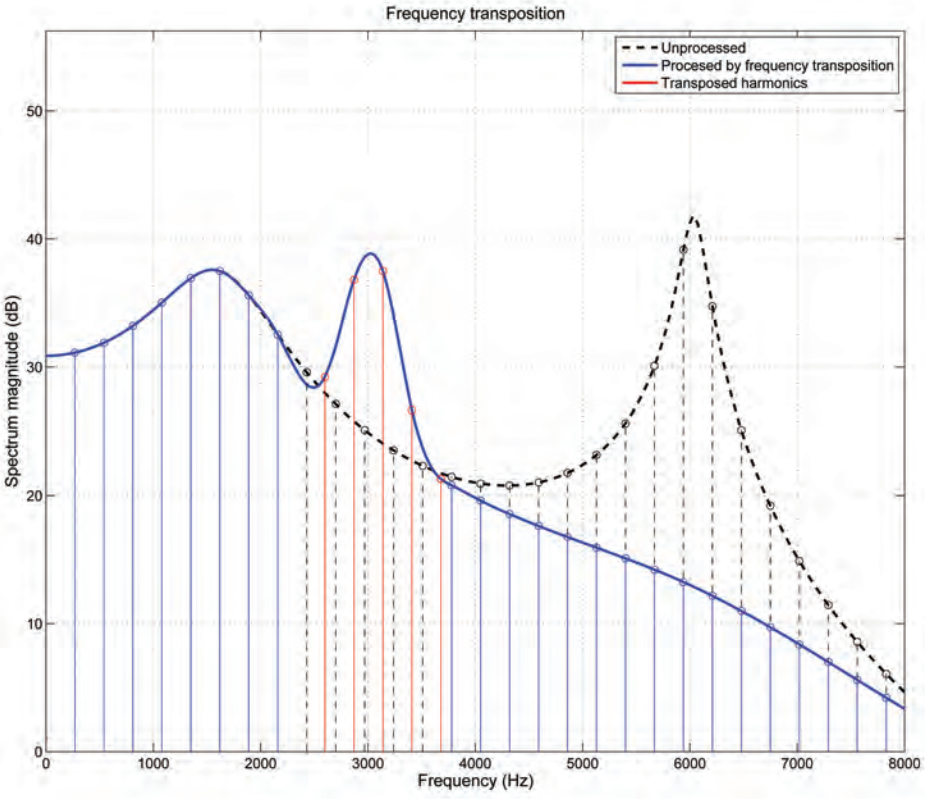


*Fig. 6. Frequency transposition moves a high frequency spectral peak to a lower frequency within the patient's range of aidable hearing. Vertical stems represent a set of harmonic frequency components, with the transposed components depicted in red. The unprocessed spectral envelope is represented by a black dashed line.*

**Frequency translation**

High frequency sounds are critical to speech intelligibility, with a substantial portion of audible speech cues occurring at frequencies higher than 3 kHz. The highest frequency speech sound, the fricative /s/, is one of the most common consonant sounds in the English language, and its energy typically peaks above 5 kHz. For some patients with high-frequency, sloping hearing loss, restoration of audibility for these high-frequency speech cues may not be possible with conventional amplification. Restoration of audibility for these patients is often constrained by the power available in the hearing aid, by the amount of gain that can be applied without intro-

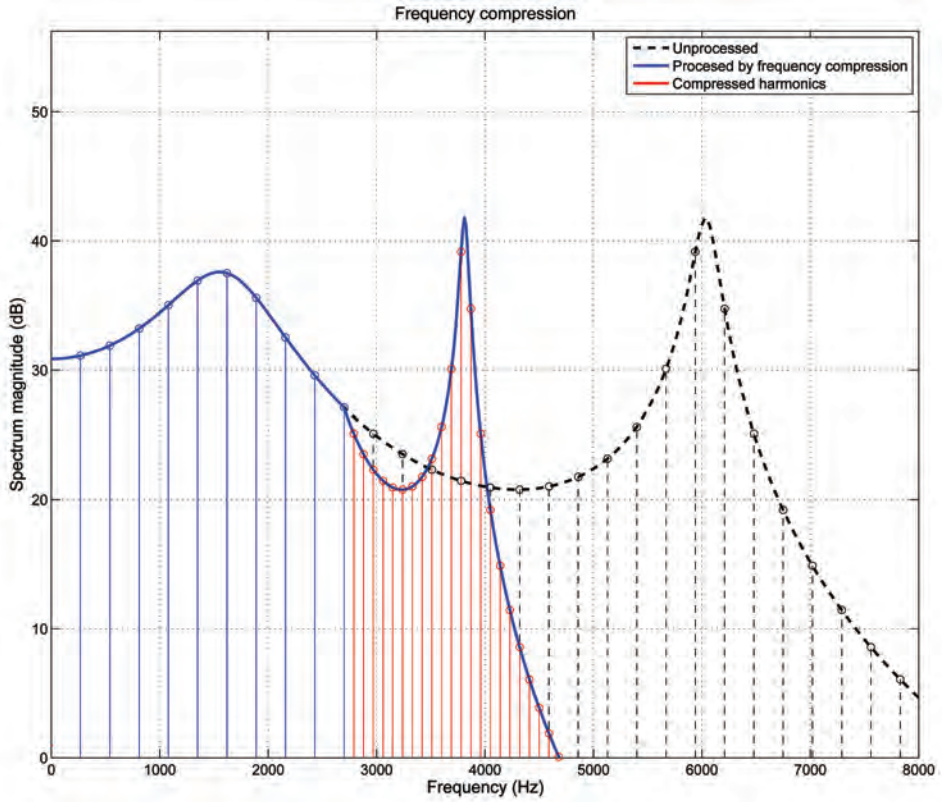


*Fig. 7. Frequency compression compresses the high-frequency part of the spectrum into a narrower frequency range. Vertical stems represent a set of harmonic frequency components, with the transposed components depicted in red. The unprocessed spectral envelope is represented by a black dashed line.*

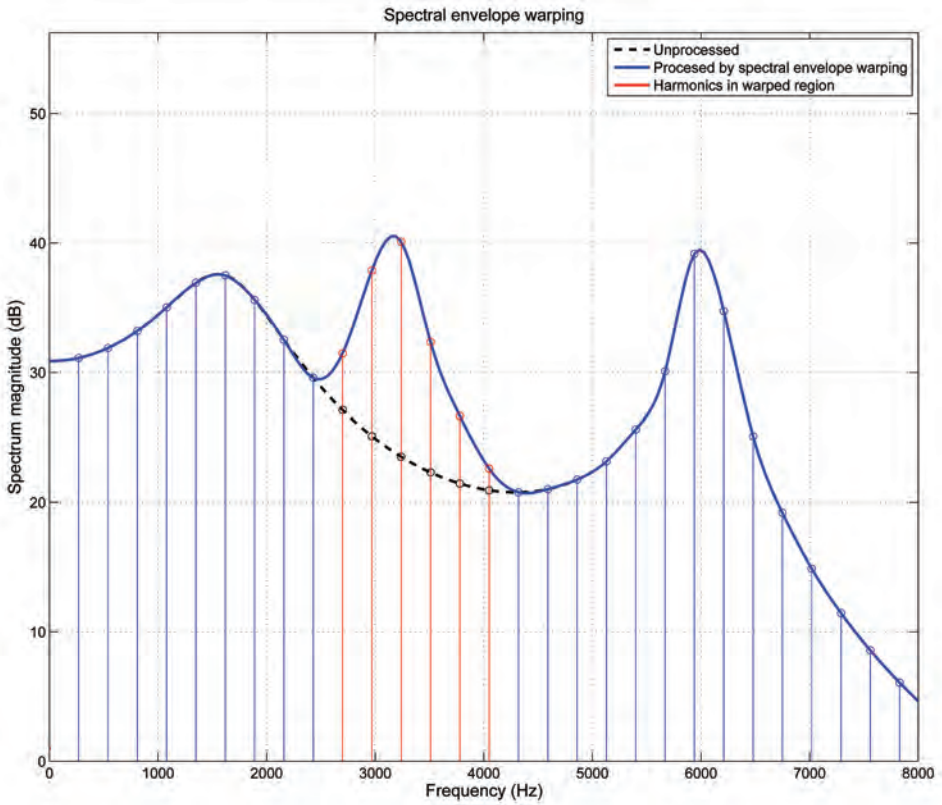


*Fig. 8. A spectral peak is detected at 6 kHz, and replicated at 3.1 kHz (red area), within the patient's aidable hearing range.*

ducing feedback oscillation, and by the extent of the cochlear damage. Even when audibility of high-frequency speech sounds can be restored, some patients with severe-to-profound hearing loss may not benefit from amplification and may perceive the amplified sound as distorted.[12]

Recent strategies for restoring audibility of these critical high-frequency speech cues have shifted the high-frequency information into lower frequency regions in which hearing loss is less severe. Such operations can be performed most easily on the frequency spectrum (which is convenient, because the above described multiband compressors operate in this domain). One strategy shifts the spectral components in the neighborhood of a high-frequency peak in the spectral envelope down to lower frequencies. This approach is described as *frequency transposition*, and is illustrated in Fig. 6 by means of an artificial frequency spectrum having two broad peaks near 1500 and 6000 Hz, the latter taken to be above the patient's range of audible hearing. The black dashed line in the spectrum plots represents the unprocessed spectrum. Frequency transposition moves the higher of the two peaks to a lower frequency, presumably within the patient's range of aidable hearing. Vertical stems represent a set of harmonic frequency components following the spectral envelope, with the transposed components depicted in red.

Another strategy warps, or compresses the high-frequency part of the spectrum into a narrower frequency range. This approach is most often called *frequency compression*, and is illustrated in Fig. 7. Here, the frequency spectrum above 2500 Hz is compressed by a factor of 3, moving the 6 kHz peak just below 4 kHz. As in the previous figure, vertical stems represent harmonic frequency components under the unprocessed spectral envelope, with components in the compressed region depicted in red.

Both frequency transposition and frequency compression risk audible distortion due to disruption of the harmonic structure of the processed sound. Harmonic components in the translated or compressed region of the spectrum do not appear at harmonic frequencies after processing. This inharmonicity is evident in the distribution of transposed and compressed harmonic components depicted in red stems in Figs. 6 and 7, respectively. These algorithms must therefore be applied with caution.

An alternative technique estimates the spectral envelope of the sound, and replicates high-frequency envelope features (peaks) at lower frequencies where they can be heard. This algorithm operates like a dynamic filter that introduces a low-frequency spectral feature whenever a high-frequency feature is detected. Because spectrum components are not translated, there is no risk of disrupting the underlying harmonic structure of the sound. This approach is illustrated in Fig. 8. Here, a peak in the spectral envelope is detected at 6 kHz and replicated at a lower frequency (3.1 kHz), within the aidable hearing range of the patient. **AT**

## References


1 S. A. Fulop, *Speech Spectrum Analysis* (Springer, Berlin, 2011).

2 T. A. C. M Claasen and W. F. G. Mecklenbräuker, "The Wigner distribution—A tool for time-frequency signal analysis, Part II: Discrete-time signals," Philips J. Res. **35**(4/5), 276–300 (1980).

3 L. Cohen, "Generalized phase-space distribution functions," J. Math. Physics **7**(5), 781–786 (1966).

4 Y. Zhao, L. E. Atlas, R. J. Marks, II, "The use of cone-shaped kernels for generalized time-frequency representations of nonstationary signals," IEEE Trans. Acoust. Speech Signal Process. **38**(7), 1084–1091 (1990).

5 D. J. Nelson, "Cross-spectral methods for processing speech," J. Acoust. Soc. Am. **110**(5), 2575–92 (2001).

6 D. J. Nelson, "Instantaneous higher order phase derivatives," Digital Sig. Proc. **12**, 416–28 (2002).

7 J. D. Markel and A. H. Gray, Jr., *Linear Prediction of Speech* (Springer, Berlin, 1976).

8 L. R. Rabiner and B.-H. Juang, *Fundamentals of Speech Recognition* (Prentice-Hall, Upper Saddle River, NJ, 1993).

9 T. F. Quatieri, *Discrete-time Speech Signal Processing* (Prentice-Hall, Upper Saddle River, NJ, 2002).

10 S. B. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," IEEE Trans. Acoust. Speech and Signal Process. **28**, 357–366 (1980).

11 B. Edwards, "Hearing aids and hearing impairment," in S. Greenberg, W. A. Ainsworth, A. N. Popper, and R. H. Fay (eds.) *Speech Processing in the Auditory System* (Springer, Berlin, 2004) pp. 339-421.

12 B. C. J. Moore, "Dead regions in the cochlea: Diagnosis, perceptual consequences, and implications for the fitting of hearing aids," Trends in Amplification **5**(1), 1–34 (2001).

Sean A. Fulop received a B.Sc. in Physics (1991) from the University of Calgary and, after M.A. degrees in Linguistics at both Calgary and UCLA, received his Ph.D. in Linguistics (1999) from UCLA. He then held temporary faculty positions in Linguistics at San José State University and the University of Chicago before joining the faculty at California State University, Fresno in 2005, where he is now Associate Professor of Linguistics and Director of Cognitive Science. His publication areas range over the fields of speech processing, phonetics, mathematical linguistic theory, and computational linguistics. His chief research programs in acoustics involve the investigation of speech sounds, and the development and dissemination of improved signal processing tools for phonetics and speech acoustics research. He is currently an Associate Editor of patent reviews for the *Journal of the Acoustical Society of America*, and has been a member of the Acoustical Society of America since 1987.

Kelly Fitz is a digital signal processing engineer specializing in the design and implementation of audio analysis, processing, and synthesis algorithms. Kelly has a Ph.D. in Electrical Engineering from the University of Illinois at Urbana-Champaign. He worked in the audio development group at the National Center for Supercomputing Applications, developing sound synthesis software for virtual reality applications, and in the Electrical Engineering department at Washington State University, where he taught signal processing and computer science, and developed algorithms for sound modeling and sound morphing. As Senior Digital Signal Processing Research Engineer at Starkey Laboratories, he conducts research combining hearing science, psychoacoustics, and signal processing to explore the perceptual consequences of hearing loss and hearing aids.

Douglas O'Shaughnessy (Massachusetts Institute of Technology, Ph.D., 1976) has been a professor at the Institut national de la recherche scientifique (INRS), University of Quebec and adjunct professor at McGill University since 1977. He is a Fellow of the Acoustical Society of America (1992) and of Institute of Electrical and Electronic Engineers (IEEE, 2006). He served 12 years as Associate Editor for the *Journal of the Acoustical Society of America*. He is the founding Editor-in-Chief of the European Association for Signal Processing (EURASIP) *Journal on Audio, Speech, and Music Processing*. He was recently elected as Vice-Chair of the IEEE Signal Processing Society (SPS) Speech and Language Technical Committee, and as member of the International Speech Communication Association (ISCA) Board, where he serves as Conference Coordinator for the series of Interspeech Conferences. He has presented tutorials on speech recognition at the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)-96, ICASSP-2001 International Conference on Communications (ICC)-2003, and at ICASSP-09. He is the author of the textbook *Speech Communications: Human and Machine* (1986 Addison-Wesley; revised 2000, IEEE Press). In 2003, with Li Deng, he co-authored the book *Speech Processing: A Dynamic and Optimization-Oriented Approach* (Marcel Dekker). He was the general Chair of ICASSP-2004.