A neurally motivated technique for voicing detection in speech.

L.S. Smith, Centre for Cognitive and Computational Neuroscience, Department of Computing Science, University of Stirling Stirling FK9 4LA, Scotland

To discover the voiced sections in (noisy) speech we sought envelope amplitude modulation in cochlear filtered speech caused by unresolved harmonics. Speech was taken to be voiced when the modulation was coherent across many channels. To achieve this, speech was Gammatone filtered, then rectified and compensated for group delay. Amplitude modulation was enhanced by applying a difference of Gaussians filter. This is effectively a bandpass filter, and the filter used was adjusted to provide a (wide) passband centred at 100Hz (210Hz) for males (females), that is at approximately the expected fundamental, F_0. AM pulses were made to stand out in each channel by separating the positive-going bandpass output and the negative-going bandpass output, and logarithmically compressing both, producing two nonnegative signals per channel, $p_i(t)$ and $n_i(t)$. An AM pulse results in a pulse in $p_i(t)$, followed by a pulse in $n_i(t)$. The $p_i(t)$ signals were summed over all the selected bands, as were the $n_i(t)$, producing $p(t)$ and $n(t)$. Bands were selected by choosing those with ERB wide enough so that F_0 harmonics might be unresolved (which depends on the centre frequency and on the sharpness (Q) of Gammatone filter). The existence of a pulse in $p(t)$ followed by a pulse in $n(t)$ was taken to signal a glottal pulse: a train of such pulses was taken to signal voicing.

Detailed results can be found in (Smith 96): we summarise here. 15 male and 15 female utterances from the TIMIT database were tested. For low noise speech, for males, varying Q between 2.3 and 9.265 makes little difference: for females, there was a significant improvement as Q was decreased from 9.265 to 2.3. In white noise (15db to 0db SNR), for both males and females, using a low value for Q considerably improves performance. The inter-glottal pulse time gives the instantaneous fundamental frequency so that the system can follow variations in F_0.

The results suggest that voicing detection would be improved if each channel was processed separately, then the voicing and F_0 found combined later. This would also remove the need to perform delay compensation.

Smith L.S. Centre for Cognitive and Computational Neuroscience Technical Report 22, CCCN, University of Stirling, July 1996.
ftp://ftp.cs.stir.ac.uk/pub/tr/cccn/TR22.ps.Z