1. In speech perception, speech signal is decomposed into a number of frequency bands. Please explain the deference between human perception（the auditory models） and Fourier transformation.

First of all, the human hearing range is commonly given as 20 to 20,000 Hz, while Fourier transformation does not put such restrict on the frequency domain.

One of the main drawbacks of Fourier transformation is that the frequency bins are linear. In contrast, the human ear responds to frequency logarithmically, not linearly, thus having different resolutions of different frequency ranges while Fourier transformation keeps a constant resolution. It is worth knowing that most of the frequencies that are of concern to us tend to be below 8 kHz. The Fourier transformation's linearity can lead to the effect that much of the Fourier transformation data is wasted on recording high-frequency information very accurately, at the expense of the low-frequency information that is generally more useful in a speech context. In respect to that, the auditory models of humans are more sensitive to lower frequencies and are suitable to be applied in the scenarios related to human voice.

There are another several powerful functions of human perception system. Like, human's auditory model can mask the sound that are not interested by human in the presence of multiple sounds, hence being able to do perception and separation simultaneously.

2. Please explain the relation of the linear spectral frequency (LSF) and the linear predictive coding (LPC).

Linear spectral frequency (LSF) uniquely represent the linear predictive coding (LPC) filter of a speech frame. Linear spectral frequencies have several properties (e.g. smaller sensitivity to quantize noise) that make them superior to direct quantization of linear predictive coding (LPC). For this reason, LSFs are very useful in speech coding. LSF's encode speech spectral information more efficiently than other transmission parameters. This can be attributed to the intimate relationship between the LSF's and the formant frequencies. Accordingly, LSF's can be quantized taking into account spectral features known to be important in perceiving speech signals. In addition, LSF's lend themselves to frame-to-frame interpolation with smooth spectral changes because of their frequency domain interpretation.

Linear predictive coding (LPC) is a method used mostly in audio signal processing and speech processing for representing the spectral envelope of a digital signal of speech in compressed form, using the information of a linear predictive model. LPC starts with the assumption that a speech signal is produced by a buzzer at the end of a tube (for voiced sounds), with occasional added hissing and popping sounds (for voiceless sounds such as sibilants and plosives). Although apparently crude, this model is actually a close approximation of the reality of speech production. LPC analyzes the speech signal by estimating the formants, removing their effects from the speech signal, and estimating the intensity and frequency of the remaining buzz. The process of removing the formants is called inverse filtering, and the remaining signal after the subtraction of the filtered modeled signal is called the residue.

3. The LPCNet uses the traditional algorithm of LPC to increase the calculation effectiveness. Do you have any idea to combine other traditional algorithms into the neural network?

Mel-Scale Frequency Cepstral Coefficients (MFCC) are widely used in various speech processing techniques, especially commonly used as features in automatic speech recognition. One key point is that the front-end processing of extracting the MFCC is considerably simple. The basic procedure to develop MFCCs is the following:

1) Convert the frequency domain from Hertz to Mel Scale

2) Take logarithm of Mel representation of audio

3) Take logarithmic magnitude and use Discrete Cosine Transformation

4) This result creates a spectrum over Mel frequencies as opposed to time, thus creating MFCCs.

Leveraging MFCCs is a fantastic way to process audio such that various Deep Learning and Machine Learning problems can learn from the recorded sounds.

The generalized cross-correlation with phase transform (GCC-PHAT) is the most popular method for estimating the time difference of arrival (TDOA) between microphones, which is an important clue for sound source localization (SSL). GCC-PHAT is computed as the inverse Fourier transform of a weighted version of the cross-power spectrum (CPS) between the signals of two microphones. The TDOA estimate is then obtained by finding the time-delay between the microphone signals which maximizes the GCC-PHAT function. Since the GCC-PHAT is a good high-level feature that represents interaural time difference, there are so many neural network based SSL systems taking the full GCC-PHAT function as the input feature and exhibiting substantial performance on DOA estimation.