How Do Computers Understand Human Speech?

Douglas O'Shaughnessy

Introduction

Alexa, Cortana, Siri, How do these commercial algorithms that interpret speech succeed in emulating human listeners? Do they actually "hear" like humans? Similarly, how do cell phones transmit sound efficiently? How do the pressure variations that constitute speech convey information? This article describes how some of these problems have been solved so that digital devices can categorize human voices. It also examines how the human voice is transformed for practical applications such as digital coding and *automatic speech recognition* (ASR). Furthermore, some devices can recognize traits of human speakers, such as identity, language, health, and emotion, and those are also outlined.

People communicate with one another by multiple means such as gestures, writing, and uttering sounds, with speech being the most efficient. At the same time, speech differs greatly from other means of communication. It consists of acoustic sounds that are only indirectly related to human concepts, and those sounds combine to create meaning to listeners who understand that specific language. Ideas in one's head create a sequence of intended words, which consist of logical speech units called *phonemes* (each language has roughly 32 of these sound units, as noted in the International Phonetic Alphabet). Muscle commands to a speaker's *vocal tract* (VT) result in movements of the tongue, lips, jaw, and velum (**Figure 1**) as air is pushed from the lungs by the diaphragm. The *vocal cords* in the larynx vibrate for most sounds (called *voiced*) at a variable rate called the *fundamental frequency* (*f_0*). In theory, an infinitely long vowel could be periodic, that is, have exact repetitions of a *pitch period,* which is the speech emitted between vocal cord closures. Such a vowel would have energy at multiples of *f_0*, called *harmonics***.**

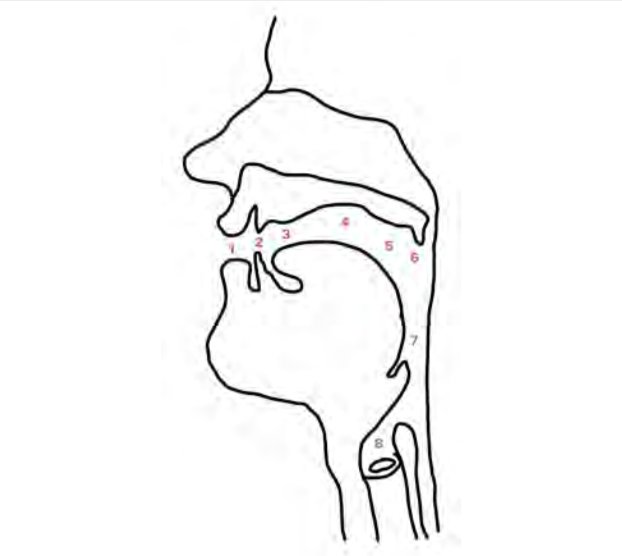


Figure 1. *Vocal tract cross section.* ***1****, Lips;* ***2****, teeth;* ***3****, alveolar ridge;* ***4****, palate;* ***5****, velum;* ***6****, uvula;* ***7****, pharynx;* ***8****, larynx.*

Air pressure variations form at points in the VT (at the vocal cords or at another constriction), with the VT acting as a filter to shape the sound waveform. When the vocal cords do not vibrate, noisy speech is called *unvoiced*. When vocal cords close abruptly, they generate energy over a broad range of perceivable frequencies. Strong voiced sounds are almost periodic and called *sonorants* (e.g., vowels), whereas weaker noisy sounds are called *obstruents*. The resulting pressure variations from the VT are speech signals, which can be deciphered by listeners or by detailed algorithms.

Artificial intelligence has been utilized to translate human speech through an algorithmic process called *analysis*, which produces a compressed version of speech for interpretation (such as conversion into text) or for efficient transmission. Analysis techniques differ across applications. In some cases, they may emulate human speech perception, whereas in other cases, they may use general mathematical models, exploiting computer power.

https://doi.org/10.1121/AT.2023.19.4.32

The value of analysis is that it can greatly reduce the information rate needed to represent speech in digital computers. For example, basic telephony transmits 64 kilobits/s by exploiting models of how the VT behaves. *Codecs* are algorithms that send speech data on digital communication channels and reconstruct the speech from the data to be understandable by listeners (while preserving naturalness), whereas automatic recognition by computer yields classifications (binary decisions for speaker verification or series of words for ASR). All these applications require data reduction because speech has much redundancy that is used for facilitating communication in adverse conditions such as noise, reverberation, and accents but that is not needed for speech analysis. For example, vowels often have eight or more repeated waveform cycles called pitch periods, but just one of those waveform cycles is enough for listeners to identify phonemes in ideal conditions.

Speech analysis uses algorithms to extract relevant information from utterances. For example, *artificial neural networks* (ANNs) are computer models of biological neural systems that allow machine perception of sight, sounds, and touch. ANNs originated decades ago (Minsky and Papert, 1969) but required modern powerful computers and "big data" to become practical. Spectral methods (e.g., Fourier transform) have also long been utilized for speech analysis (Fourier, 1822). Early artificial intelligence used human-designed "expert" systems (Reddy et al., 1973), now replaced by fully automatic systems. Recent *end-to-end* ANNs do all speech analysis by automatic learning based on observed data. For decades, *hidden Markov models* (HMMs) dominated the speech recognition field. A HMM is a statistical model with states that represent probabilities for data spectra during sequential sections of speech, with transitions between states that model the variable timing of speech (Rabiner, 1989).

Some signal processing can apply to a diverse range of data (audio, video, other physical measurements). However, speech is different from other signals. Speech presents a unique challenge to signal processing because of its highly encoded, dynamic nature. Thus, correlating speech with its meaning using analysis techniques is far more complex than a simpler process like classifying objects in images. To understand speech analysis choices, let us first examine human speech communication.

What Is Speech Communication?

Communication via speech involves its production (coding) and perception (decoding). In artificial communication (e.g., radio, Morse code, sign language), coder and decoder are directly related and may be inverse operations. In speech, coding and decoding differ greatly. Over the course of evolution, mammals evolved to have similar auditory mechanisms to survive, which facilitate hearing time and frequency patterns in sounds but are not necessarily specific for speech. Early mammal communications were likely simple bursts of noise, where breaths were interrupted by constrictions in the VT (Lieberman, 1984). Nowadays, human VTs emit sounds that are easy to interpret (Fitch, 2000). Ordinary breathing can create noise if the vocal cords are narrowed. The 0- to 4-kHz range is the most useful for perception, having approximately 1 resonance/kHz (Fant, 1970). Speech energy tends to decrease with frequency due to the low-pass nature of puffs of air from the glottis.

Speech consists of sequences of words in utterances that are organized by the rules of syntax and semantics that people internalize when learning language. Each spoken word is made up of a series of *phones,* which are physical sounds for intended phonemes. Phones vary in periodicity, intensity, and spectrum; speech analysis exploits these domains.

In creating dynamic speech, speakers vary the VT shape to create sounds with different *formants*, which are resonances whose center frequencies are acoustic cues to listeners to distinguish different phones. In addition to conveying the identities of phonemes to listeners, speech also has information about syntactic structure and emphasis, conveyed by *intonation.* This term covers a range of acoustic phenomena that include f_0, sound amplitude, and phone durations, whose relationships to linguistic information are highly complex.

Speech waveforms vary greatly in time (**Figure 2**), even for the same words produced by one speaker. This is partly due to many small variations in spectral phase, which result from VT losses (friction, thermal) and minute airflow variations. These variations do not convey useful phonetic information, and thus speech analysis often is designed to discard phase as a useful cue. Speakers instead articulate to achieve VT shapes that yield suitable spectral amplitudes (formant frequencies),

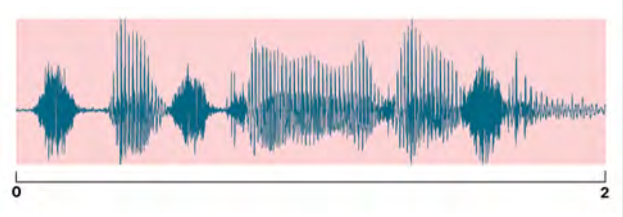


Figure 2. *Speech waveform of the utterance "Speech communication." Note the six strong vowel portions interspersed with weaker consonants.* ***Nos. on bottom:*** *time in seconds.*

and listeners focus on this intended detail for deciphering speech.

Precisely which aspects of speech are relevant for human communication are difficult to demonstrate. Controlled perceptual experiments with human listeners use synthetic stimuli emulating speech (Liberman, 1957). These demonstrate that certain changes in physical stimuli can reliably evoke phonetic perception, but there is only circumstantial evidence that these are actually employed in speech. Because it is difficult to prove direct relationships between acoustic features and perception, some neural speech algorithms avoid all preprocessing (analysis) of speech, instead training recognition models directly on speech waveforms (Ravanelli and Bengio, 2018). However, such neural models effectively learn processing in a way that is similar to mainstream speech analysis.

Analog-to-Digital Conversion

Speech signals vary continuously in time, and so it becomes important that temporal details in speech be preserved in the analysis process. Digital computers need sequences of data for processing, so the *analog-to-digital conversion* process takes sample values from signal intensity at uniformly spaced intervals. The interval at which the computer sampling occurs is called the sample or *Nyquist* rate. This rate must exceed twice the highest frequency in the signal (Picone, 1993) to allow proper representation of the temporal features of signals. Furthermore, an analog low-pass filter must first suppress all higher frequency energy to prevent distortion in the reconstructed speech of codecs. Given the bandwidth (typically 300-3,200 Hz) of standard telephony, 8,000 samples/s (hertz) is a commonly used sampling rate. Nontelephone applications are wideband because listeners may discern frequencies up to 20 kHz (Jacewicz et al., 2023). Some applications (such as compact discs) have sample rates of 44.1 kHz. Other common applications include internet audio at 16 kHz.

Dynamic Energy over Limited Time Ranges

Because speech is nonstationary (dynamic), characteristic measures vary in time. Typically, analysis averages measures over brief time ranges called windows, repositioned regularly at a periodic frame rate. A common standard for much of speech processing is 100 frames/s (Spanias, 1994) that accommodates *coarticulation,* which is VT motion from phone to phone (Öhman, 1966). Speech averages approximately 12 phones/s and has both anticipatory and lagging effects of VT organ movements. The simplest relevant measure of speech is its energy. Energy can help distinguish classes (e.g., vowels from fricatives) as well as distinguish speech from background noise.

Periodicity

Besides energy, the other most salient feature of voiced speech is periodicity. Vocal cords vibrate around 100 times/s for men and 200 times/s for women. The physical f_0 is heard as perceptual *pitch,* as the brain processes the timing and locations of auditory neural firings along the basilar membrane of the inner ear (Moore, 1995). The f_0 is useful in tone languages to distinguish words phonemically and in most languages, syntactically and semantically to delimit phrasal units, distinguish yes/no questions from statements, and give emphasis (O'Shaughnessy, 1979).

Estimation of the f_0 (Rabiner et al., 1976) exploits the regularity of vocal cord closures, each of which causes a speech energy increase, with ensuing gradual decay until the next closure (**Figure 2**). Waveform peaks related to strong harmonics in the first formant often confound simple f_0 estimation. Also, "periods" in sonorants have small deviations in amplitude (*jitter*) and timing (*shimmer*) (Horii, 1979). Most f_0 detection algorithms search for peaks in either the speech waveform or its spectrum and assume small changes from period-to-period (except at voiced/unvoiced transitions).

Measuring periods is most reliable if the signals are simplified in spectral amplitude and phase. Most f_0 detectors do this by reducing the spectral detail irrelevant to periodic structure (spectrum flattening). A process called *autocorrelation* yields a zero-phase and squared-amplitude spectrum, convolution of a signal with its time-reversed version. This measure, which eliminates phase while retaining spectral amplitude, is also widely used in telephony codecs.

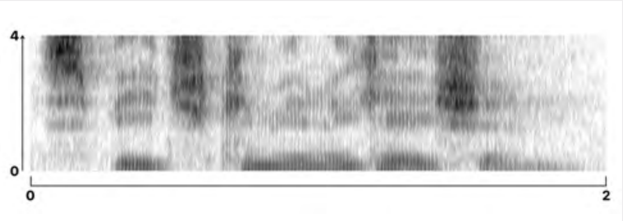
Spectral Analysis

The features of energy and f_0 help classify speech versus nonspeech (Rabiner and Sambur, 1975), but most applications require much more information about the speech signal. A discrete Fourier transform (DFT) provides an energy representation of speech, consisting of N spectral samples (N being the window duration; Picone, 1993). N can be as small as 10 for codecs to model 4-5 formants (using 2 parameters to represent each resonance) or as many as hundreds (when seeking details over multiple pitch periods). Various analysis methods represent the spectral distribution of speech energy because this information correlates well with many aspects of speech production and perception (Fant, 1970).

VT shapes for basic vowels of most languages (/i, a, u/) have widely spaced formants, as seen in spectral displays (**Figure 3**). For consonants, concentrations of energy vary consistently with *place of articulation* (VT constriction). Relevant communicative cues are found in the energy peaks, not in the valleys.

Speech codecs and ASR often use a version of DFT called *subband coding* (SBC) (Crochiere et al., 1976) that exploits the greater energy and better resolution found in human speech and hearing at lower frequencies. In SBC, speech is separated into M distinct spectral ranges by band-pass filters, with ranges following the perceptual *Mel scale*. In this scaling, spacing is linear below 1 kHz and logarithmically wider above 1 kHz; hence, there is less precision as frequency increases (Davis and Mermelstein, 1980). This scale reflects the distribution of sensory hair cells along the basilar membrane. Instead of a DFT transforming N time samples of speech into N spectral values, SBC yields much reduced M time signals of smaller bandwidths, which allows better exploitation of the distribution of spectral information. For codecs, each filter output (*channel*) uses smaller step sizes for more precision at (more perceptually useful) lower frequencies. ASR, which need not preserve waveform detail, simply calculates M channel energies. Some older vocoders used about 20 channels (Picone, 1993); modern ANN ASR uses $M = 40–100$ (Mohamed et al., 2022).

Figure 3. *Wideband speech spectrogram (utterance in* ***Figure 2****). The darkness displays logarithmic energy. Time in seconds is on the* x*-axis and frequency in hertz is on the* y*-axis. Formants show as bands that are roughly horizontal but vary in time.*



A major challenge for speech analysis is to efficiently represent phonetic information in each frame of speech. One choice is to create hundreds of DFT spectral samples versus 10 *linear predictive coding* (LPC) coefficients (see **Linear Predictive Coding**) (Makhoul, 1973). In most speech (sonorants), the focus is to model major aspects of a spectrum of 4-5 resonances, which appear as a modulation superimposed on dozens of harmonics. High-bit-rate codecs often directly replicate speech samples and minimize a success criterion called the *signal-to-noise ratio* (noise from quantization), but other applications seek phonetic data at frame rates (100/s) much lower than the sampling rates (8,000/s).

Some popular applications (e.g., MP3 players) use a direct encoding of speech spectra in *adaptive transform coding* (Zelinski and Noll, 1977). Although SBC uses a small set of filters, ATC retains all spectral samples. ATC must then inform its decoder about dynamic quantizer parameters. These step sizes and numbers of bits for all spectral samples are assigned in proportion to rough estimates of energy.

A final spectral measure for speech is the *zero-crossing rate*, which simply counts the times the waveform changes the algebraic sign (baseline 0 is normal atmospheric pressure in silence). It roughly estimates the dominant frequency in speech, being low for sonorants and high for noisy sounds. Combined with energy, it can be used for discriminating speech from background noise.

Time Windows

Speech is dynamic. It has phones of finite durations, changing center frequencies and bandwidths of resonances and varying phase. When viewing sonorants through windows, to observe these varying phenomena, DFT does not show (theoretical) discrete components because harmonics are spread over a frequency range inversely proportional to window duration. Typical *narrowband* spectrograms use a window with multiple periods (bandwidth less than f_0 to visualize harmonics), whereas *wideband* spectrograms display precise timing transitions and clear formants (**Figure 3**).

Most speech applications have more interest in the broader envelope of spectra than harmonics because the former relates to VT shape, whereas the latter varies with excitation. Both are relevant for speech coding and synthesis, as their outputs require full signals for human listeners, but ASR focuses on the VT shape for phone identification, frequently excluding consideration of excitation (it is also difficult to integrate relevant information over different timescales).

Linear Predictive Coding

Speech has temporal correlations at widely different ranges: local (within pitch periods), midrange (coarticulation across phones), and global (syntactic and semantic aspects across words). These variations in temporal complexity during the speech production process complicate the speech analysis process compared with signals simpler than speech. For instance, high bit rate codecs using logarithmic quantizers accommodate the non-uniform probability distribution of speech sample amplitudes but do not exploit the relevant temporal variations of speech.

Most speech analysis exploits short time features of speech, which correspond to the VT shape and spectral envelope. One analysis technique is to compare the difference between each speech waveform sample and an estimate of that sample, based on N immediately prior samples (**Figure 4**). This difference is usually far smaller than the differences between the samples themselves, thus allowing smaller step sizes and reduced quantization noise.

The predicted estimate used is typically a linear combination of $N = 10$ prior samples. Codecs in cellular telephony still use this traditional LPC (Makhoul, 1973). As sonorants have primary excitation in each period at vocal cord closure, ensuing samples largely follow the *impulse response* of the VT (signal from the VT filter with one-sample input).

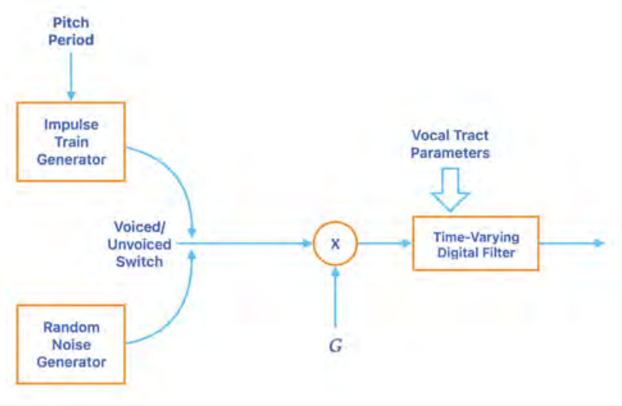


Figure 4. *Linear predictive synthesizer. Noise or periodic pulses (at the fundamental frequency* (f_0) *rate) simulating breaths from the lungs are multiplied (X) by a gain factor (G). The filter amplifies the resonance frequencies as determined by the vocal tract shape.*

Mel-Frequency Cepstral Coefficients

For decades, a common method of speech analysis has been mel-frequency cepstral cefficients (MFCCs) (Davis and Mermelstein, 1980). Cepstral analysis originated for *deconvolution,* estimating both components of a filtering. For example, output speech $s(n)$ is modeled as coming from a filter with VT impulse response $h(n)$ excited by input $e(n)$, which is constriction noise or glottal puffs. Cepstral analysis can be used for dereverberation, radar/sonar, and speech. Speech is often viewed as periodic or noisy input to a VT filter; thus, cepstral analysis can estimate both filter and its excitation input.

Although MFCCs are common for speech analysis, simpler logarithmic *band-pass filter energies* (BFEs) (MFCC, but without the final inverse transform) are increasingly used. The inverse step yields uncorrelated parameters, but it does not correspond to human perception. To get each $c(n)$ value, one multiplies the mel-deformed spectrum by an n-period sinusoid and then averages. MFCCs are used because the combined information of all $c(n)$ represents enough detail to distinguish phones.

MFCCs or BFEs can represent the static position of the VT in each frame of speech, but VT velocity and acceleration are also useful measures (Picone, 1993). Thus, static MFCCs are often augmented by 13 *delta* (frame difference) values and 13 *delta-delta* values. Using 13 MFCCs, one can discriminate spectral differences of approximately 100 Hz, which corresponds to small differences in formants in languages with many vowels, such as English. For example, tongue height differences correlate to f_1 in the range of 300-700 Hz for 5 vowels (/i, I, e, E, ae/).

Analysis methods suffer when audio has distortions (e.g., environmental/channel noise or reverberation). Current methods treat speech spectra globally without distinguishing perceptually important prominences. Future analysis could focus on spectral peaks because they are salient amid typical distortions. Such approaches have been avoided in modern ASR due to the difficulty of integrating such information with common frame-based methods.

Artificial Neural Networks

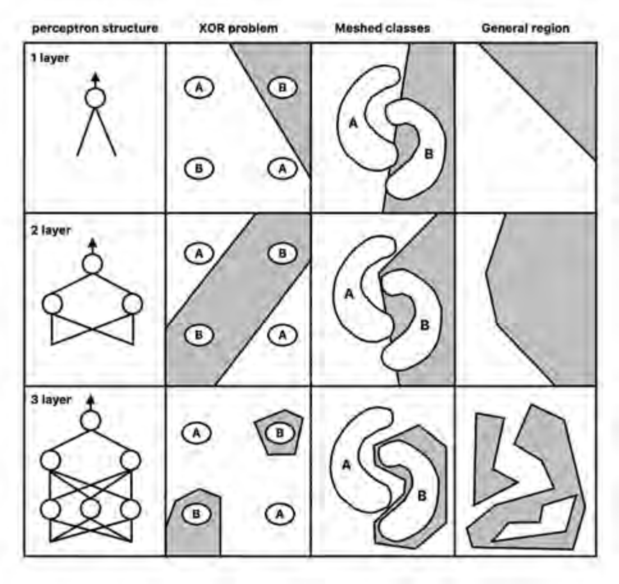
Now we discuss the major tool that is currently used to process almost all speech applications. The function of an ANN is to convert an input data sequence to an output sequence. By using a huge number of nonlinear operations, an ANN has potentially excellent processing power. ANNs are trained on large amounts of data, guided by a *cost* or *loss function* to minimize entropy or a mean-square difference between a target and estimated signals. For ASR or speaker verification, network inputs are speech samples or frame-based spectral representations (MFCCs or BFEs), and the outputs can be probabilities for text corresponding to the speech or a decision on speaker identity. For applications such as speech coding and speech enhancement, an encoding ANN outputs compressed data (for transmission), and then a decoding ANN maps back to reconstructed speech samples (or corresponding spectral vectors).

Because ANNs are based on natural neural systems, consider a biological neuron. Its output is binary (a brief pulse known as an action potential, when the weighted sum of its inputs exceeds a specified threshold). In the human nervous system, neurons at the initial processing group or *layer* receive sensory information (from the eyes and ears). Nodes in an ANN use a nonlinear threshold operation (*activation function*) that is generally smooth and monotonic to facilitate mathematical differentiation, which allows the use of derivatives for gradient-descent parameter modification in iterative training (where the model parameters are updated in proportion to the slope of a *loss* function).

In the simplest form of ANN, each layer has nodes feeding outputs to the next layer. Biological neural networks have hundreds of billions of neurons, whereas ANNs usually have millions of nodes. Nodes may have operations other than binary nonlinear weighting, such as *pooling* (selecting a maximal value among inputs) (Scherer et al., 2010).

For classification of input data, one can visualize an *N*-dimensional representation space, where *N* is the number of samples, such as a speech waveform (or spectral) sequence. Each node with these *N* inputs then determines a flat surface in the space by the linear combination of its weighted inputs; 0/1 output specifies either side (**Figure 5**). Each node thus can act as an elementary

Figure 5. *Possible regions for multi-layer artificial neural networks (ANNs). The shaded/unshaded regions, bordered by* ***black lines****, are two estimated regions for a classification problem between objects* ***A*** *and* ***B*** *(two classes whose borders are shown with* ***circles*** *or* ***lines****). More layers allow more complex regions. ©1987 IEEE. Figure reprinted, with permission, from R. Lippmann, (1987), "An introduction to computing with neural nets."* IEEE ASSP Magazine*, 4(2), 4-22.*



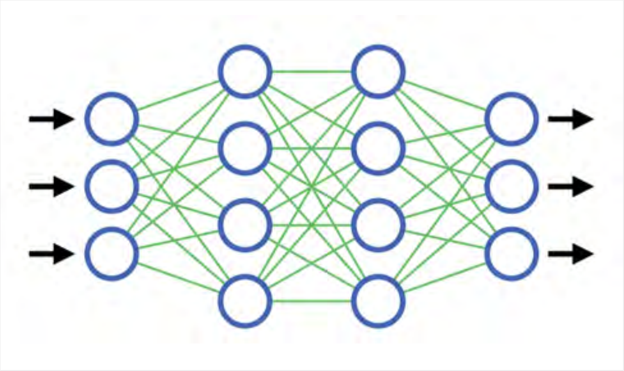


Figure 6. *Structure of a simple 4-layer ANN.* ***Circles:*** *nodes in vertical sets (layers) connected by links (**green lines**) where flow progresses from **left to right**. Each link corresponds to a weighting, and each node sums the weighted inputs and outputs as a thresholded binary value.*

classifier. With three or more layers, an ANN may optimize the combined locations of surface boundaries of complex class regions in the space. In addition, choices for model parameters allow for decisions that are more complex than binary. Such complexity is often needed to handle the huge variability in many applications, including speech. However, this complexity hinders heuristic interpretation of ANNs. The parameters of ANNs (node weights and biases) are available during design, but the ANNs complex operation greatly hinders direct parameter manipulation (debugging).

ANN parameters are trained to minimize a differentiable loss function, which is modeled as a cost to minimize errors. Direct minimization of errors is infeasible because the relationship between network parameters and classification errors is extremely complex. As ANNs derive directly from many examples, they must avoid *overfitting,* where models become too close to matching observed data points when using limited training. To generalize models, training data are often modified by artificial distortion (additive noise and/or deletion of random portions in time and frequency) (Ko et al., 2015).

Basic ANNs are *fully connected feedforward neural networks* (FFNNs), meaning that all nodes in each layer feed all nodes in each successive layer (**Figure 6**). This, however, is overly general for most applications because patterns to be analyzed tend to have a diversity of local and global aspects. For example, objects often occupy only small portions of an image or identifying a vowel using BFEs may only need small subsets (limited frequency or temporal portions of a spoken vowel). To exploit the often-local nature of classification, one may use *convolutional neural networks* (CNNs) (LeCun and Bengio, 1995). A CNN processes input data over small ranges called *receptive fields.* CNNs were first developed for image recognition, to enhance edges. Applied to speech, CNNs can filter formants in a spectrogram. However, edges in spectrograms are less relevant as features for speech than they are for images.

Whereas CNNs exploit local data correlations, *recurrent* networks handle longer range patterns (Schuster and Paliwal, 1997). Pertinent information in speech is distributed very unevenly in time and frequency. Thus, sections of speech with low energy are far less useful than portions with strong formants, and coarticulation and intonation affect speech over tens and hundreds of frames, respectively. Both ANNs and HMMs struggle to exploit this nonuniform distribution of information; basic ANNs do best with static patterns (Lippmann, 1987).

Recurrent neural networks (RNNs) have architectures with feedback to get over the problem of uneven distributions of variability, such as those found in speech. They use distributed hidden states that store information about past input. A common recurrent method is *long short-term memory* (Hochreiter and Schmidhuber, 1997). In human perception, listeners internalize portions of speech (several seconds) in some analyzed form in their short-term memory. Utilizing such a wide range of data in FFNNs and CNNs is exceedingly difficult. The range of analysis of an RNN can extend well beyond the very limited scope of CNN kernels or of context-dependent HMMs.

Automatic Speaker Verification

Automatic speaker verification (ASV) is a speech-analysis task that has followed research like ASR, despite being a very different task. ASR extracts phonetic information from VT shape via acoustic analysis, whereas ASV distinguishes different VTs. ASV can be more difficult to accomplish than other speech tasks because what distinguishes behavioral output such as speech is far less definitive. Impostors can simulate others' voices, and recordings can be used surreptitiously (*spoofing*) (Wu et al., 2015).

Final Comments

This article has considered how to analyze speech to understand how humans and machines go about their perception. Spectrograms formed the basis of speech analysis until 1970. A major breakthrough in the speech analysis and decoding field was LPC, which is still used in cell phones today. Versions of spectral analysis have been used for speech applications, including SBC and MFCCs. Although ANNs have existed for 50 years, they are only recently dominating applications for speech, due to improvements in computing power and the availability of large databases.

Because many of the analysis methods were developed years ago, one may speculate about future breakthroughs. Efficient methods have not yet come close to human performance for many speech applications, and current approaches are fragile. For example, ASR trained on limited data does not generalize well to variations in speakers, contexts, and environmental degradations such as noise. Human listeners can handle the huge variability of speech from different speakers and under many distortions. ANNs try to generalize via various types of regularization, but such methods do not reflect many actual acoustic conditions. Also, current methods struggle to exploit the full range of information in speech, given the diverse ways that phonetics, syntax, and semantics are embedded. Anyone who has struggled with Siri or Alexa devices to understand what they are trying to say can relate. Hence, the speech-analysis field has more to explore.

Acknowledgments

This work was supported by the Natural Sciences and Engineering Research Council (NSERC) of Canada Grant 142610.

References

Crochiere, R. E., Webber, S. A., and Flanagan, J. L. (1976). Digital coding of speech in sub-bands. *Bell System Technical Journal* 55, 1069-1085. https://doi.org/10.1002/j.1538-7305.1976.tb02929.x.

Davis, S., and Mermelstein, P. (1980). Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions on Acoustics, Speech, and Signal Processing* 28, 357-366. https://doi.org/10.1109/TASSP.1980.1163420.

Fant, G. (1970). *Acoustic Theory of Speech Production*. Walter de Gruyter, Berlin, Germany.

Fitch, W. T. (2000). The evolution of speech: A comparative review. *Trends in Cognitive Sciences* 4, 258-267. https://doi.org/10.1016/S1364-6613(00)01494-7.

Fourier, J. B. (1822). *Théorie Analytique de la Chaleur*. Cambridge University Press, Cambridge, UK.

Hochreiter, S., and Schmidhuber, J. (1997). Long short-term memory. *Neural Computation* 9, 1735-1780. https://doi.org/10.1162/neco.1997.9.8.1735.

Horii, Y. (1979). Fundamental frequency perturbation observed in sustained phonation. *Journal of Speech, Language, and Hearing Research* 22(1), 5-19. https://doi.org/10.1044/jshr.2201.05.

Jacewicz, E., Alexander, J. M., and Fox, R. A. (2023). Extended high frequency in hearing and speech. *Acoustics Today* 19(3), 22-29.

Ko, T., Peddinti, V., Povey, D., and Khudanpur, S. (2015). Audio augmentation for speech recognition. In *Proceedings of the 16th Annual Conference of the International Speech Communication Association* (INTERSPEECH 2015). Dresden, Germany, September 6-10, 2015, pp. 3586-3589.

LeCun, Y., and Bengio, Y. (1995). Convolutional networks for images, speech, and time series. In Arbib, M. A. (Ed.), *The Handbook of Brain Theory and Neural Networks*. MIT Press, Cambridge, MA, pp. 3361-3374.

Liberman, A. M. (1957). Some results of research on speech perception. *The Journal of the Acoustical Society of America* 29(1), 117-123. https://doi.org/10.1121/1.1908635.

Lieberman, P. (1984). *The Biology and Evolution of Language*. Harvard University Press, Cambridge, MA.

Lippmann, R. (1987). An introduction to computing with neural nets. *IEEE ASSP Magazine*, 4(2), 4-22.

Makhoul, J. (1973). Spectral analysis of speech by linear prediction, *IEEE Transactions on Audio and Electroacoustics* 21, 140-148. https://doi.org/10.1109/TAU.1973.1162470.

Minsky, M., and Papert, S. (1969). *Perceptrons*. MIT Press, Cambridge, MA.

Mohamed, A., Lee, H. Y., Borgholt, L., Havtorn, J. D., et al. (2022). Self-supervised speech representation learning: A review. *IEEE Journal of Selected Topics in Signal Processing* 16, 1179-1210. https://doi.org/10.1109/JSTSP.2022.3207050.

Moore, B. C. (Ed.). (1995). *Hearing*. Academic Press, Cambridge, MA.

Öhman, S. E. (1966). Coarticulation in VCV utterances: Spectrographic measurements. *The Journal of the Acoustical Society of America* 39(1), 151-168. https://doi.org/10.1121/1.1909864.

O'Shaughnessy, D. (1979). Linguistic features in fundamental frequency patterns. *Journal of Phonetics* 72, 119-145. https://doi.org/10.1016/S0095-4470(19)31045-9.

Picone, J. W. (1993). Signal modeling techniques in speech recognition. *Proceedings of the IEEE* 81(9), 1215-1247. https://doi.org/10.1109/5.237532.

Rabiner, L. R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE* 77(2), 257-286. https://doi.org/10.1109/5.18626.

Rabiner, L. R., and Sambur, M. R. (1975). An algorithm for determining the endpoints of isolated utterances. *Bell System Technical Journal* 54, 297-315. https://doi.org/10.1002/j.1538-7305.1975.tb02840.x.

Rabiner, L., Cheng, M., Rosenberg, A., and McGonegal, C. (1976). A comparative performance study of several pitch detection algorithms. *IEEE Transactions on Acoustics, Speech, and Signal Processing* 24(5), 399-418. https://doi.org/10.1109/TASSP.1976.1162846.

Ravanelli, M., and Bengio, Y. (2018). *Speech and Speaker Recognition from Raw Waveform with SincNet*. arXiv:1812.05920. https://do 10.48550/arXiv.1812.05920.

Reddy, D., Erman, L., and Neely, R. (1973). A model and a system for machine recognition of speech. *IEEE Transactions on Audio and Electroacoustics* 21, 229-238. https://doi.org/10.1109/TAU.1973.1162456.

Scherer, D., Müller, A., and Behnke, S. (2010). Evaluation of pooling operations in convolutional architectures for object recognition. In Diamantaras, K., Duch, W., and Iliadis, L.S. (Eds.), *Artificial Neural Networks—ICANN 2010*: 20th International Conference on Artificial Neural Networks, Thessaloniki, Greece, September 15-18, 2010, pp. 92-101.

Schuster, M., and Paliwal, K. K. (1997). Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing* 45, 2673-2681. https://doi.org/10.1109/78.650093.

Spanias, A. S. (1994). Speech coding: A tutorial review. *Proceedings of the IEEE* 82(10), 1541-1582. https://doi.org/10.1109/5.326413.

Wu, Z., Evans, N., Kinnunen, T., Yamagishi, J., Alegre, F., and Li, H. (2015). Spoofing and countermeasures for speaker verification: A survey. *Speech Communication* 66, 130-153. https://doi.org/10.1016/j.specom.2014.10.005.

Zelinski, R., and Noll, P., (1977). Adaptive transform coding of speech signals. *IEEE Transactions on Acoustics, Speech, and Signal Processing* 25(4), 299-309. https://doi.org/10.1109/TASSP.1977.1162974.

Contact Information

Douglas O'Shaughnessy
douglas.oshaughnessy@inrs.ca

National Institute of Scientific Research (INRS)
Place Bonaventure
Montreal, Quebec H5A 1K6, Canada

For author bio, please go to acousticstoday.org/bios-19-4-3