Machine Based Decoding of Voices and Human Speech

Alessandro Vinciarelli

1 Introduction

Figure 1 shows the *Threshold of Hearing* - the minimum intensity required for a sound to be heard - as a function of frequency. The lowest part of the curve corresponds to the frequencies typical of human speech (roughly between 20 and 400 Hz). Thus, human ears are most sensitive to human voices than to any other sound in the environment. From an evolutionary point of view, the most likely explanation is that speech has been a key advantage for our species [7]. Therefore, it is not surprising to observe that technology has made major efforts aimed at dealing automatically with speech signals [19].

The earliest technological approaches revolving around speech signals date back to the first half of the Twentieth century. It is in this period that the diffusion of telecommunications has fostered the development of *coding systems*, i.e., of automatic methodologies capable to represent a signal in a form as compact as possible. The main goal of these efforts was to improve the efficiency of transmissions, i.e., to convey as much information as possible using as little data as possible. In parallel, research laboratories started to work on *Automatic Speech Recognition* (ASR), the task of automatically transcribing speech signals [19]. After a pioneering stage, it is during the seventies that ASR technologies make the most important progress. The reason is twofold. On the one hand, it is in such a period that computers have become powerful enough to deal with the ASR problem. On the other hand, it is in the seventies that the statistical methodologies still today underlying most ASR approaches make their first appearance in the speech technology community.

The initial ASR attempts targeted relatively simple tasks like the automatic transcription of phone numbers. In this case, the predefined list of words that a speech recognition system can actually transcribe is limited ("*zero*", "*one*", "*two*", ..., "*nine*"). Furthermore, the utterances are *not connected*, i.e., they are separated by silences long enough to easily segment the speech stream into individual words. Over the years, the efforts have addressed increasingly more challenging recognition tasks. First the automatic transcription of people reading written texts (the data does not include noise, there are no disfluencies or grammatical errors, language models can constrain effectively the space of the transcription hypotheses to be searched) then the recognition of spontaneous

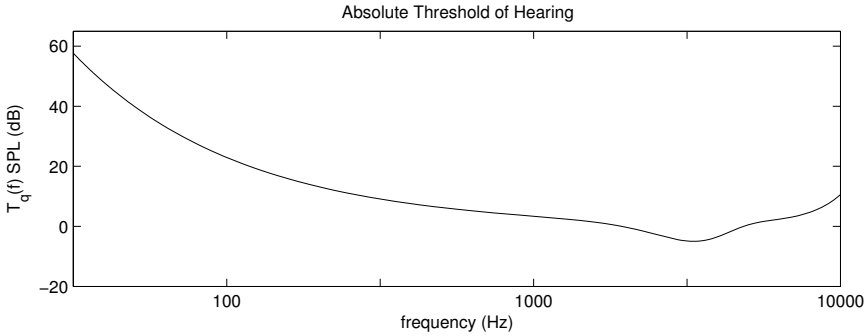

Figure 1: Absolute Threshold Of Hearing (TOH). The TOH is plotted on a logarithmic scale and shows how the energy necessary to hear frequencies between 50 and 4000 kHz is significantly lower than the energy needed for other frequencies.

speech in naturalistic settings (the data includes noise, there are disfluencies and grammatical errors, the language models constrain to a limited extent the space of possible transcription hypotheses). Applications like *Siri* and *Cortana*, capable to effectively interact with their users via speech, come at the end of this long process and rely on unprecedented large volumes of data available through the development of Internet based services and the diffusion of mobile platforms.

Typically, ASR approaches include a *normalisation* step aimed at eliminating, or at least attenuating, variability in the speech signal that is not relevant to the automatic transcription problem. Normalisation methodologies target the suppression of variability due to sources like, e.g., echoes or environmental noise. Furthermore, they target nonverbal and paralinguistic aspects of speech like, e.g., prosody (loudness, pitch, speaking rate, etc.), vocalisations (laughter, crying, etc.), use of silence and pauses, overlapping speech, turn-taking, etc. The reason is that these elements do not change the transcription (*what people say*) even if they contribute to its sense (*how people say it*). However, the last $10 - 15$ years have witnessed increasingly more efforts aimed at analysis and understanding of nonverbal components of speech, especially in fields like *Computational Paralinguistics* [23] and *Social Signal Processing* (SSP) [31, 32]. This has led to approaches for the automatic analysis of a wide spectrum of social and psychological phenomena that speech conveys, including emotions, personality, dominance, roles, effectiveness of delivery, etc. The efforts in this direction have made it clear that it is not possible to correctly transcribe speech without taking into account communicative aspects that paralanguage and nonverbal communication convey. However, a full integration between ASR and SSP has still to be achieved.

The goal of this chapter is to provide a short introduction to the technologies mentioned above, in particular when it comes to the main technological and

methodological issues and components. The rest of the chapter is organised as follows: Section 2 introduces Automatic Speech Recognition and its state-of-the-art, Section 4 shows how the computing community deals with nonverbal aspects of speech and Section 4 draws some conclusions.

2 Automatic Speech Recognition

ASR is the task of automatically transcribing speech data. In mathematical terms, this corresponds to map a signal $S = (s_0, s_1, \ldots, s_N)$ into a sequence of words $W = (w_1, w_2, \ldots, w_T)$, where s_k is the k^{th} sample of the signal and N is the total number of samples in S. Sample s_k is a physical measurement - typically air pressure - made at time $k\Delta t$, where Δt is the length in seconds of the time interval between two consecutive measurements. When using a microphone, the physical measurement that accounts for air pressure is the displacement of an elastic membrane, positioned inside the microphone, with respect to its position of equilibrium. The value of Δt is constant during a recording and it is called *sampling period*. Its inverse is the *sampling frequency* F in Hertz, i.e., the number of times per second that a measurement has been done during the recording. In the case of speech, the typical sampling frequency is 44 kHz when high quality is required (e.g., broadcast material or commercial audio products) and 8 kHz when low quality is sufficient (e.g., phone and radio communications).

Figure 2 shows the main components of an ASR system. The *front-end* is the step that takes the signal S as input and gives as output a representation of it suitable for further processing. In current state-of-the-art ASR technologies, the representation is a sequence $X = (\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_M)$ of observation vectors, where M is the total number of vectors in sequence X. The observation vector $\vec{x}_k$ is extracted from a short analysis window - the typical length is 30 ms - that starts at time kB, where B is the interval of time between the start of two consecutive analysis windows (in the most frequent case, $B = 10$ ms). Typically, two consecutive windows are partially overlapping (with the parameters mentioned above the overlapping is 20 ms).

The rationale behind such a representation is that spoken sentences are sequences of *phonemes*, the atomic sounds that compose every word in a given language. Ideally, there should be one observation vector per phoneme, but it is not possible to know a-priori where the phonemes are. Windows positioned at regular time steps do not correspond exactly to phonemes. However, they are expected to entirely include one phoneme at least in some cases. This is the reason why the windows must be long enough to frequently enclose one phoneme, but short enough to rarely include two consecutive phonemes. The use of statistical approaches for the transcription step (see below) allows one to deal with the uncertainty in the position of the windows with respect to the actual phonemes.

The second stage, the actual transcription step, takes X as input and gives as output the sequence of words $W = (w_1, \ldots, w_T)$ that is the final output of

the ASR system. In general, the transcription relies not only on X, but also on two linguistic resources, namely the *lexicon* and the *language model*. The lexicon L is the list of words that the system can actually give as output. In other words, every $w_i \in W$ must be one of the entries of the lexicon L. If the signal contains a word that is not in the lexicon, the system will still give as output one of the words of the lexicon, typically the one that is closest from a phonetic point of view. The language model is a probability distribution $p(W)$ that estimates how probable a given transcription W is. Language models are typically obtained by counting the occurrences of individual words and N-grams (sequences of N consecutive words) in large corpora of text. The main role of lexicon and language model is to constrain the space of the hypotheses to search, i.e., to eliminate those transcriptions that are too unlikely to be considered.

State-of-the-art ASR systems find the transcription $\hat{W}$ that satisfies the following equation:

$$\hat{W} = \arg \max_{W \in \mathcal{W}_L} p(X, W)p(W) \tag{1}$$

where $p(X, W)$ is a probability distribution defined over the joint space of observation and word sequences and $\mathcal{W}_L$ is the set of all possible sequences of words belonging to the lexicon L. In other words, an ASR system takes into account all possible transcriptions for a given observation sequence X and, for each transcription, estimates the probability $p(X, W)p(W)$. Then, the transcription that corresponds to the largest probability is retained as the actual transcription of the input speech data. Given that the number of transcriptions is prohibitively large, lexicon and language model are used to eliminate all transcriptions that are unlikely to match the observation sequence X.

The description above shows that, from a technical point of view, the most distinctive aspects of an ASR system are the type of information that X conveys and the approach adopted to estimate $p(W, X)p(W)$. Furthermore, the lexicon L typically characterises the application domain for which the ASR system has been designed, from the simple transcription of phone numbers (only ten items in the lexicon corresponding to the digits from zero to nine) to the transcription of unconstrained conversations (up to $100,000$ items in the lexicon expected to cover $90 - 95\%$ of all words used in a generic conversation). The rest of this section focuses on front-end and automatic transcription.

2.1 Front-End

The extraction of the observation vectors from the speech signal is typically referred to as *feature extraction* because the components of the observation vectors are called *features*. These latter are physical measurements expected to convey information relevant to the recognition of the words being uttered. In particular, the features are expected to be different and stable for different phonemes. The assumption of stability over time intervals comparable to the duration of a phoneme is known as *piecewise quasi-stationarity assumption* and underlies virtually every ASR approach proposed in the literature.

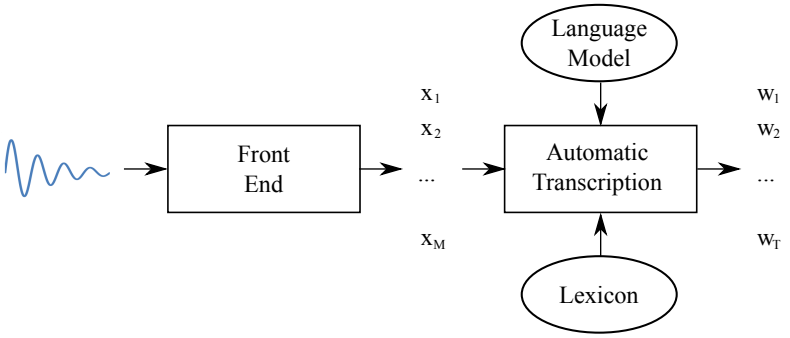

Figure 2: The Figure shows the main technological components of an Automatic Speech Recognition System. The front-end takes as input the speech signals and gives as output a sequence of observation vectors. The Automatic Transcription maps the sequence of vectors into a sequence of words.

The features most commonly extracted from the speech signals are the *Mel-Frequency Cepstral Coefficients* (MFCC) [35] and the *Perceptual Linear Prediction* coefficients (PLP) [12]. In both cases, the goal is to obtain a smooth version of the *spectral envelope*, i.e., the curve of the frequency-amplitude plan that describes the way the energy of a sound is distributed across different frequencies. The main difference between the actual distribution and the envelop is that this latter is designed to be steady (no jumps of the first derivative) and smooth (no major oscillations) while following as close as possible the actual distribution. In intuitive terms, the spectral envelope can be thought of as the curve that connects the maxima of a spectrum, hence the use of the term *envelop*.

MFCC and PLP coefficients are the most commonly adopted features, but the literature provides a large number of other methodologies. However, the overall attempt is always to account for the spectral properties in the most compact possible way while conveying all the information necessary to correctly transcribe the signals.

2.2 Automatic Transcription

The transcription step aims at finding the sequence of lexicon words $\hat{W}$ that satisfies Equation (1). In intuitive terms, $\hat{W}$ is the sequence of lexicon entries that maximises the joint probability of W and X multiplied by the probability of W. The main approaches adopted in the literature to estimate $p(X, W)$ and $p(W)$ are the *Hidden Markov Models* [20] and the *N-gram models* [17].

The main assumption underlying Hidden Markov Models (HMM) is that there is a sequence of non-observable (hidden) states underlying the sequence of observations X. In the case of ASR, the sequence of the states corresponds to a sequence of phonemes that compose the words in W. Typically, there are three states for every phoneme, namely *onset*, *apex* and *offset*. From a technical

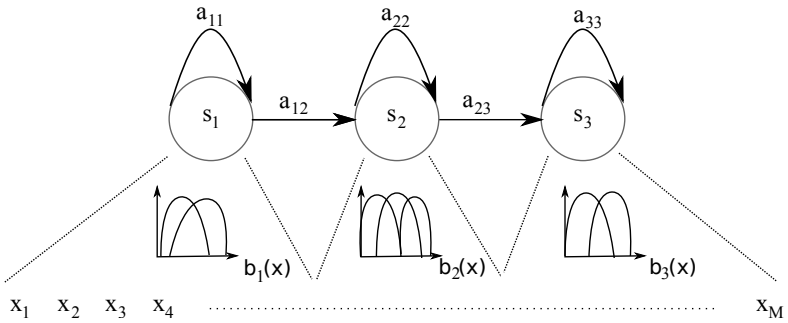

Figure 3: The figure shows how a left-right HMM works. The emission probability density functions associated to the states allow one to estimate the probabilities of an observation vector $\vec{x}$ belonging to one of the states, the transition probabilities allow one to estimate the probability of passing from one state to the other.

point of view, the sequence of states corresponding to a word is obtained by concatenating multiple HMMs each corresponding to a phoneme. Overall, the expression of $p(X, W)$ in the case of a HMM is as follows:

$$p(X, W) = \pi_{s_1} b_{s_1}(\vec{x}_1) \prod_{k=2}^{M} a_{s_k s_{k-1}} b_{s_k}(\vec{x}_k) \tag{2}$$

where s_j is the j^th state in the sequence of states that underlies W, π_{s_1} is the probability of starting with state s_1 (i.e., the probability of starting the sequence with a certain phoneme), $a_{s_k s_{k-1}}$ is the probability of a transition between state s_{k-1} and s_k, and $b_{s_k}(\vec{x})$ is the probability of observing $\vec{x}$ when the underlying state is s_k (the *emission probability function*). Since self-transitions are possible, the HMMs can accommodate variations in length of the same phoneme (multiple observations can be attributed to the same underlying state).

The values of π_{s_k} are typically obtained by counting the number of times in a given collection of spoken data an utterance starts with a certain phoneme and, hence, the underlying HMM starts with a certain state. Similarly, the transition probabilities are estimated by counting how frequently in a collection of spoken data a given state s_i is followed by another state s_j. In the case of the emission probability functions, the most common approach to get the explicit expression of $b_s(\vec{x})$ is the application of the Expectation-Maximization [3]. In general, the emission probability functions correspond to Mixtures of Gaussians:

$$b_s(\vec{x}) = \sum_{k=1}^{G} \alpha_k \mathcal{N}(\vec{x}|\Sigma_{ks}, \vec{\mu}_{ks}) \tag{3}$$

where the *mixing coefficients* α_k sum up to 1, $\mathcal{N}(.)$ is a multivariate Gaussian, Σ_{ks} and $\vec{\mu}_{ks}$ are covariance matrix and mean of Gaussian k in the mixture of state s. Figure 3 shows how the HMMs work in practice.

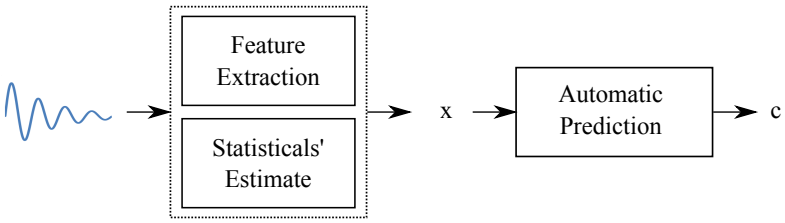

Figure 4: The Figure shows the main technological components of an Automatic Speech Recognition System. The front-end takes as input the speech signals and gives as output a sequence of observation vectors. The Automatic Transcription maps the sequence of vectors into a sequence of words.

The transcription step actually consists in finding the sequence of states (hence of phonemes and words) that better accounts for the observation sequence X. Such a task is performed with the *Viterbi Algorithm* [9] that is capable to find the sequence of states that maximises the probability $p(W, X)$. However, the search through all possible sequences W can be constrained with a language model $p(W)$ so that the computational effort is reduced. The most common approach to estimate $p(W)$ is the N-gram model. The reason is not only that these models appear to be the most effective, but also that they naturally fit the Viterbi algorithm. The expression of $p(W)$ with an N-gram model is as follows:

$$p(W) = \prod_{k=N}^{T} p(w_k | w_{k-1}, w_{k-2}, \ldots, w_{k-N+1}) \tag{4}$$

where N is the *order* of the model, w_k is the k^{th} word of W and T is the total number of words in W. While being simple, the N-gram models have been shown to be more effective than models trying to take into account the meaning of the words or the grammar of a language.

The main role of $p(W)$ in the transcription is to lower the probability of transcriptions that, while making sense from an acoustic point of view, are not necessarily observable in a given language. This applies in particular to short sequences that include only *function words*, i.e., terms that are content-independent, but allow one to build grammatically correct sentences (articles, prepositions, etc.). Such short sequences include expressions such as "*there is a*", "*it is on*", etc., that, while being frequent, often can be confused with longer words. In general, content words are less frequent than function words (roughly one third of all the words appearing in a corpus occur only once), but the acoustic evidence is sufficiently strong to counterbalance the low value of $p(W)$.

3 Nonverbal Vocal Behaviour

Besides extracting features from the speech signal, the front-end of an ASR system typically tries to eliminate any source of variablity that is not relevant to the automatic transcription of what is being said, a step typically referred to as *normalisation*. This applies to variability resulting from gender, age, emotional state, accent, speaking style and any other factor that, while influencing the way something is said, does not influence the transcription of an utterance. The reason is that the ASR performances increase when there is a consistent relationship between the phonemes being uttered and the features being extracted. However, these sources of variability have recently become the focus of domains like Computational Paralinguistics [23] and Social Signal Processing [32]. The reason is that nonverbal components of speech - prosody, voice quality, vocalisations, disfluencies, intonation, etc. - convey socially and psychologically relevant information about speakers and their interaction.

Figure 4 shows the main technological components of systems that analyse nonverbal communication in speech. Like in the case of ASR, speech data is first segmented into short analysis windows (typically $20 - 30$ ms) that overlap each other and start at regular time-steps (typically 10 ms). In this way, it is possible to extract short-term properties from the speech signals, i.e., properties that can be expected to be relatively stable for no more than a few tens of milliseconds or the time that someone can hold a stable configuration of the articulators. The result is that, for a given short-term property, it is possible to obtain as many measurements as there are analysis windows that fit in the data (in the case of 30 ms windows that start at regular time-steps of 10 ms, one second of speech yields 970 measurements). These measurements are then summarised with statisticals like average, variance, entropy, minimum, maximum, etc. In this way, while not taking into account every single value of a measurement, it is still possible to have an idea of its distribution over a speech sample.

The ultimate goal of the process above is to represent a speech sample as a vector of physical properties that account for how a person speaks. Once such a vector is available, it is then possible to apply statistical approaches that can be used to infer the traits attributed by people. The short-term properties adopted in the different works presented in the literature cover the most important speech properties. The measurements most commonly adopted target pitch (the fundamental frequency), energy and speaking rate, i.e., the *Big-Three* of prosody. Intuitively, these measurements account for the sound of the voice, how loud a person speaks and how fast she does it, respectively. The statisticals account for the distribution of the measurements. In particular, the average accounts for the values that occur most frequently, the variance for how wide is the range of the measurement, the entropy for its variability, minimum and maximum (used only rarely because they can be outliers) for the dynamic range, etc. Other short-term properties account for voice spectral properties such as Mel Frequency Cepstral Coefficients (MFCC) [35], Harmonic-to-Noise ratio [4], spectral tilt [14], etc.

The prediction step is performed using a wide spectrum of classification and

regression approaches. The former are adopted when nonverbal behaviour is adopted to infer categorical information like, e.g., which of the six basic emotions a speaker is displaying [22] or which is the role that someone is playing in a meeting [10]. The latter are adopted when nonverbal behaviour is used to infer dimensional information like, e.g., emotions represented in the Valence-Arousal space or personality assessed along the Big-Five traits. From a mathematical point of view, a classifier is a function $f(\vec{x})$ that maps a vector $\vec{x}$ into c, where this latter is a class that belongs to a predefined set $\mathcal{C} = \{c_1, c_2, \ldots, c_N\}$ (N is the total number of classes). In contrast, a regressor is a function $f(\vec{x})$ is a function that maps a vector $\vec{x}$ into a real number y. The choice between classifiers and regressors depends on the particular problem being targeted. The literature proposes a large number of classifiers and regressors, but the most popular and effective are, e.g., Support Vector Machines [11], Deep Neural Networks [2], LASSO [28], Logistic Regression [13], Gaussian Processes [21], etc.

While in the case of ASR all systems address the same problem - the automatic transcription of speech recordings - in the case of Social Signal Processing and Computational Paralinguistics, the technologies presented in the literature address a large number of different issues. The earliest approaches focused on emotion (see [22] for an extensive survey), but the latest technologies have applied methodologies like those described above to infer from speech information such as role, conflict, dominance, synchrony, interest, personality, developmental disease in children, depression, etc. The rest of this chapter provides a short state-of-the-art of the main problems addressed in the literature.

3.1 Analysis of Social Signals and Paralanguage

The computing literature proposes a large number of approaches aimed at inferring socially and psychologically relevant phenomena from speech recordings [23, 31, 32]. One of the problems that have been addressed most extensively is emotion recognition, i.e., the automatic identification of the emotions that speakers experience based on the physical characteristics of their speech. The problem has been the subject not only of many articles (see [22] for an extensive survey), but also of several international benchmarking campaigns in which a large number of different approaches have been adopted and compared [24, 25]. As a result, it has been possible to perform a meta-analysis showing that there is a set of features - called the *Geneva Minimalistic Acoustic Parameter Set* (GeMAPS) - that appear to be more reliable than the others in conveying information about the emotional state of a speaker [8].

In recent years, the attention has shifted towards other problems that can be addressed using the approach depicted in Figure 4. In particular, a large number of works and an international benchmarking campaign [26] have been dedicated to the inference of personality traits - both self-assessed and attributed - from nonverbal aspects of speech [30]. No feature set has been shown to be more reliable than others like in the case of the emotions. However, a few indications emerge from the literature. The first is that all works addressing the problem of the relationship between speech and personality from a computing point of view

adopt trait-based personality models, i.e., models that represent personality as a D-dimensional vector where every component accounts for a behavioural dimension. In most cases, the traits correspond to the *Big-Five*, five major dimensions that have been shown to capture most individual differences (Openness, Conscientiousness, Extraversion, Agreeableness, Neuroticism) [5]. The second is that the only task that can be performed with satisfactory performance is the prediction of whether a person is above or below median with respect to every trait. Finally, the third is that not all traits can be inferred equally well from speech. In particular, the performances tend to be satisfactory only for Extraversion and Conscientiousness.

Another domain where there has been a significant effort has been the recognition of the roles that people play in a particular social setting [6, 10, 34]. The main difference with respect to the problems mentioned above is that, in this case, it is necessary to analyse recordings that include multiple voices. However, the proposed approaches remain similar to those depicted in Figure 4. The only difference is that there is a preliminary speaker diarisation step, i.e., a segmentation of the audio recordings into interval in which only one person is expected to speak [1, 29]. This allows one to use not only the features that have been adopted for the inference of emotions and personality, but also features that account for turn-taking, including speaking time distribution across speakers, adjacency matrices, amount of overlapping speech, etc. [31, 32]. The main limitation of this area is that the roles tend to be specific of a given setting and, unlike the case of constructs like personality or phenomena like the emotions, it is not possible to identify a general set of roles that applies to all possible contexts.

In the case of other social and psychological phenomena, the number of works was not sufficiently large to give rise to a research community, but the approaches stem directly from those adopted for the other problems mentioned so far in this section and, overall, replicate the scheme of Figure 4. Such phenomena include, e.g., dominance [15], conflict [16], mimicry [18], depression [27], interest [33], etc.

4 Conclusions

This chapter has provided a general introduction to the problem of machine based decoding of speech and human voice. Such a definition encompasses all technologies that automatically infer information from speech signals, whether this means to automatically transcribe what a speaker is saying - the domain is, in this case, Automatic Speech Recognition - or to predict information about social and psychological aspects of speakers and their interactions with others - the domains are, in this case, Social Signal Processing and Computational Paralinguistics.

One of the main messages of the chapter is that the approaches adopted to infer information from speech can be described in terms of two general schemes. The first, underlying most Automatic Speech Recognition systems, has been de-

picted in Figure 2. The second, underlying most systems aimed at Social Signal Processing and Computational Paralinguistics, has been depicted in Figure 4. A crucial problem in both cases is the extraction of features, i.e., automatic measurements that account for the physical properties of speech. In the case of the problems that have been addressed most extensively in the literature (ASR and emotion recognition), it has been possible to identify feature sets that are more reliable than the others or, at least, lead to satisfactory performances in the majority of the experimental settings. In the case of those problems that have been addressed more recently and less extensively, the identification of reliable features is still an open issue.

For what concerns the machine intelligence aspect, i.e., the computational approach aimed at mapping the features into information of interest (transcriptions or social and psychological phenomena), the state-of-the-art in ASR is the adoption of Hidden Markov Models, statistical sequential models that can take into account temporal aspects of the data they take as input. In the case of SSP and Computational Paralinguistics, the variety of approaches is wider because the data is typically represented with a single vector and, then, any type of classifier or regressor can be applied. However, Deep Neural Networks have started to be used more and more frequently both in ASR and in the other domains and they are likely to become one of the most common approach, if not the dominant approach, in the next years.

The main application field of the technologies described in this chapter is likely to be Human-Computer Interaction in all its multiple aspects, from speech based personal assistants like Siri and Cortana, to social robots expected to understand the inner state of their users. The main challenges concern the possibility to work in naturalistic environments where noise and lack of constraints make it difficult to extract proper features and to constrain the space of possible outcomes, respectively. Furthermore, speech technologies are used increasingly more frequently in non-technological fields such as, e.g., social psychology and cognitive neuropsychology. This will hopefully result into new insights about human speech and its crucial role in our life.

References

[1] X. Anguera, S. Bozonnet, N. Evans, C. Fredouille, G. Friedland, and O. Vinyals. Speaker diarization: A review of recent research. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(2):356–370, 2012.

[2] Y. Bengio. Learning deep architectures for AI. *Foundations and Trends in Machine Learning*, 2(1):1–127, 2009.

[3] J. Bilmes. A gentle tutorial of the EM algorithm and its application to parameter estimation for gaussian mixture and hidden markov models. Technical Report 510, International Computer Science Institute, 1998.

[4] G. de Krom. A cepstrum-based technique for determining a harmonics-to-noise ratio in speech signals. *Journal of Speech, Language, and Hearing Research*, 36(2):254–266, 1993.

[5] J.M. Digman. The curious history of the Five-Factor Model. In J.S. Wiggins, editor, *The Five-Factor Model of Personality*, pages 1–20. Guilford Press, 1996.

[6] W. Dong, B. Lepri, A. Cappelletti, A. Pentland, F. Pianesi, and M. Zancanaro. Using the influence model to recognize functional roles in meetings. In *Proceedings of the International Conference on Multimodal interfaces*, pages 271–278, 2007.

[7] R. Dunbar, L. Barrett, and J. Lycett. *Evolutionary psychology: A beginner's guide*. Oneworld Publications, 2005.

[8] F. Eyben, K.R. Scherer, B. W Schuller, J. Sundberg, E. André, C. Busso, L. Y Devillers, J. Epps, P. Laukka, and S.S. Narayanan. The geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing. *IEEE Transactions on Affective Computing*, 7(2):190–202, 2016.

[9] G.D. Forney. The Viterbi algorithm. *Proceedings of the IEEE*, 61(3):268–278, 1973.

[10] N.P. Garg, S. Favre, H. Salamin, D. Hakkani Tür, and A. Vinciarelli. Role recognition for meeting participants: an approach based on lexical information and social network analysis. In *Proceedings of the ACM International Conference on Multimedia*, pages 693–696. ACM, 2008.

[11] M.A. Hearst, S.T. Dumais, E. Osman, J. Platt, and B. Scholkopf. Support Vector Machines. *IEEE Intelligent Systems and their Applications*, 13(4):18–28, 1998.

[12] H. Hermansky. Perceptual linear predictive (PLP) analysis of speech. *Journal of the Acoustical Society of America*, 87(4):1738–1752, 1990.

[13] D.W. Hosmer Jr and S. Lemeshow. *Applied logistic regression*. John Wiley & Sons, 2004.

[14] P. Jackson, M.and Ladefoged, M. Huffman, and N. Antoñanzas Barroso. Measures of spectral tilt. *Journal of the Acoustical Society of America*, 77(S1):S86–S86, 1985.

[15] D.B. Jayagopi, H. Hung, C. Yeo, and D. Gatica-Perez. Modeling dominance in group conversations using nonverbal activity cues. *IEEE Transactions on Audio, Speech, and Language Processing*, 17(3):501–513, 2009.

[16] S. Kim, F. Valente, M. Filippone, and A. Vinciarelli. Predicting continuous conflict perceptionwith Bayesian Gaussian processes. *IEEE Transactions on Affective Computing*, 5(2):187–200, 2014.

[17] C.D. Manning and H. Schütze. *Foundations of statistical natural language processing*. MIT Press, 1999.

[18] S. Michelet, K. Karp, E. Delaherche, C. Achard, and M. Chetouani. Automatic imitation assessment in interaction. In *International Workshop on Human Behavior Understanding*, pages 161–173, 2012.

[19] R. Pieraccini. *The voice in the machine*. MIT Press, 2012.

[20] L.R. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.

[21] C.E. Rasmussen. *Gaussian processes for machine learning*. MIT Press, 2006.

[22] K.R. Scherer. Vocal communication of emotion: A review of research paradigms. *Speech Communication*, 40(1):227–256, 2003.

[23] B. Schuller and A. Batliner. *Computational paralinguistics: emotion, affect and personality in speech and language processing*. John Wiley & Sons, 2013.

[24] B. Schuller, A. Batliner, S. Steidl, and D. Seppi. Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge. *Speech Communication*, 53(9):1062–1087, 2011.

[25] B. Schuller, S. Steidl, and A. Batliner. The INTERSPEECH 2009 emotion challenge. In *Proceedings of Interspeech*, pages 312–315, 2009.

[26] B. Schuller, S. Steidl, A. Batliner, E. Nöth, A Vinciarelli, F. Burkhardt, R. Van Son, F. Weninger, F. Eyben, and T. Bocklet. The INTERSPEECH 2012 speaker trait challenge. In *INTERSPEECH*, volume 2012, pages 254–257, 2012.

[27] F. Scibelli, A. Troncone, L. Likforman-Sulem, A. Vinciarelli, and A. Esposito. How major depressive disorder affects the ability to decode multimodal dynamic emotional stimuli. *Frontiers in ICT*, 3:16, 2016.

[28] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society B*, pages 267–288, 1996.

[29] S.E. Tranter and D.A. Reynolds. An overview of automatic speaker diarization systems. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(5):1557–1565, 2006.

[30] A. Vinciarelli and G. Mohammadi. A survey of personality computing. *IEEE Transactions on Affective Computing*, 5(3):273–291, 2014.

[31] A. Vinciarelli, M. Pantic, and H. Bourlard. Social Signal Processing: Survey of an emerging domain. *Image and Vision Computing Journal*, 27(12):1743–1759, 2009.

[32] A. Vinciarelli, M. Pantic, D. Heylen, C. Pelachaud, I. Poggi, F. D'Errico, and M. Schroeder. Bridging the Gap Between Social Animal and Unsocial Machine: A Survey of Social Signal Processing. *IEEE Transactions on Affective Computing*, 3(1):69–87, 2012.

[33] M. Yeasin, B. Bullot, and R. Sharma. Recognition of facial expressions and measurement of levels of interest from video. *IEEE Transactions on Multimedia*, 8(3):500–508, 2006.

[34] M. Zancanaro, B. Lepri, and F. Pianesi. Automatic detection of group functional roles in face to face interactions. In *Proceedings of the International Conference on Multimodal Interfaces*, pages 28–34, 2006.

[35] F. Zheng, G. Zhang, and Z. Song. Comparison of different implementations of MFCC. *Journal of Computer Science and Technology*, 16(6):582–589, 2001.