Speech Perception by Humans and Machines

Matthew H. Davis[1]

Odette Scharenborg[2]

(1) Medical Research Council: Cognition and Brain Sciences Unit, Cambridge, UK

(2) Centre for Language Studies, Radboud University, Nijmegen, the Netherlands

Correspondence to: Matt Davis
Medical Research Council, Cognition & Brain Sciences Unit
15 Chaucer Road
Cambridge
CB2 7EF
UK

Email: matt.davis@mrc-cbu.cam.ac.uk

Telephone: +44 1223 273 637

Abstract:

Humans are more effective than machines at recognizing speech. This advantage for human listeners is particularly pronounced for speech that is heard against background noise, contains unfamiliar words or is degraded in other ways. Yet, automatic speech recognition (ASR) systems have made substantial advances over the past few decades and are now in everyday use by millions of people around the world. In this chapter we provide a brief explanation of how ASR systems operate. We then suggest three ways in which these systems could potentially be improved by capitalising on knowledge of human speech recognition.

Acknowledgements:

Matt Davis was supported by the UK Medical Research Council (MC-A060-5PQ80), and Odette Scharenborg was supported by a Vidi-grant from the Netherlands Organization for Scientific Research (NWO; grant number: 276-89-003).

In 2009 a company called SpinVox was the subject of media controversy after it was revealed that the voicemail transcription service that it supplied was dependent on call centres in South Africa and the Philippines. Rather than the fully automated system that some had anticipated, behind the scenes they used human listeners to transcribe many or perhaps all the voicemail messages that they processed. In 2010 SpinVox was sold to a computer speech technology company, Nuance Communications, who had previously acknowledged that "Spinvox is offering something that is impossible to deliver now"[1]. At the time of writing it remains unclear whether automated transcription of voicemail messages – from any speaker, on any topic, and with the background noise and distortion that is common in telephone calls – will ever achieve the level of accuracy of human listeners.

Simply put, the most effective system for perceiving speech and recognizing words is a human who is a native speaker of the target language and has intact hearing. This advantage for human listeners is particularly pronounced for speech that is heard against background noise, contains unfamiliar words or is degraded in other ways. The goal of the other chapters in this volume is to understand how human listeners achieve this remarkable success. This is curiosity-driven science at its most vital and informative. Despite the rise of other means of communication such as e-mail and messaging services on smartphones, spoken language remains the primary form of human communication. The cognitive and neural processes that support successful spoken communication are unique to humans and in many ways define what it is that makes us human (Pinker, 1994).

[1] John West from Nuance's mobile group quoted at: http://www.techweekeurope.co.uk/networks/voip/spinvox-faked-speech-transcription-service-and-broke-privacy-1451

Knowledge of how the human brain perceives and understands speech also has more pragmatic purposes, which are the focus of this chapter. Our focus here is on linking insights from human speech perception to help listeners that are not human, i.e. computer speech recognition systems. This could be considered a key technological application of research on human speech perception and spoken word recognition[2] – however, in practice engineering approaches to automatic speech recognition have been (at best) only loosely guided by knowledge gained from studying human speech perception. Indeed, perhaps the most famous comment on this topic comes from the pioneer of Automatic Speech Recognition (ASR) systems, Fred Jelinek who apparently remarked in the 1980s "Anytime a linguist leaves the group the recognition rate goes up" (see Jurafsky and Martin, 2009). The development of machine speech recognition systems has proceeded in isolation from the study of human speech recognition. A goal of this chapter is to attempt to bridge this divide – both by explaining the operation of current state-of-the-art machine recognition systems to researchers studying human speech recognition, and by highlighting mechanisms that allow human listeners to achieve their remarkable success in speech comprehension that are potentially useful for ASR systems. While Jelinek was perhaps right to dismiss linguists in favour of engineers at the time, we believe that our current understanding of human speech perception can offer useful insights to engineers building automatic speech recognition systems.

In this chapter we will first report on the current status of the recognition of speech by machines before describing the underlying computations by which current ASR systems operate. We will then consider three ways in which insights from

[2] Two further applications of research in human speech perception are to help listeners who are hearing impaired (see Mattys et al., 2012) or are not native speakers (see Chen & Marian, this volume).

human speech recognition might guide future technological advances in machine speech recognition. These proposals entail three different forms of human-inspired design in which: (1) the nature of the representations, (2) the computational implementation or (3) the functions achieved during recognition are modelled on human speech perception. Specifically, we seek inspiration from human recognition by: (1) adopting articulatory feature representations modelled after the mechanics of human speech production, (2) using brain-inspired processing mechanisms (deep neural networks, DNNs), (3) incorporating forms of perceptual learning that appear to operate in human listeners.

1. Current status of machine speech recognition

Many of us already use speech recognition technology such as Apple's Siri, Google Now or Microsoft Cortano on a daily basis when interacting with our smartphones. These systems are practical, and extremely effective. However, at present none of these systems reach 100% accuracy in transcribing single sentences. Such suboptimal recognition performance is particularly noticeable in large vocabulary ASR systems that have to deal with a wide variety of speakers, degraded speech signals, or different types of background noise. In his seminal 1997 paper, Lippmann showed that machines perform more than an order of magnitude worse than humans on a word recognition task in degraded conditions (Lippmann, 1997). But despite a large improvement of machine performance in noisy or degraded conditions in recent years, automatic systems still perform 6-7 times worse than humans (e.g., Hilger & Ney, 2006 on similar material as used for the comparison by Lippmann, 1997).

Scharenborg (2007) reviewed the results of systematic comparisons of human and machine recognition systems and documented an order-of-magnitude better performance for humans, not only at the word-level (Lippman, 1997; Carey & Quang, 2005; Juneja, 2012), but also at the level of individual phonemes (e.g., Cutler & Robinson, 1992; Meyer et al., 2006; Sroka and Braida, 2005), and at the level of articulatory/acoustic features (e.g., Cooke, 2006; Sroka and Braida, 2005; Meyer et al., 2011). This difference in performance persists even if higher-level lexical, semantic, or world knowledge is prevented from influencing perception. This is shown by a detailed comparison of human and machine performance on a corpus of logatomes, i.e., CVC and VCV sequences without semantic information (e.g., Meyer et al., 2006). Thus it is not the case that human advantages at speech recognition are solely due to more effective comprehension and use of higher-level linguistic information. Artificial systems are impaired at perception as well as comprehension of speech.

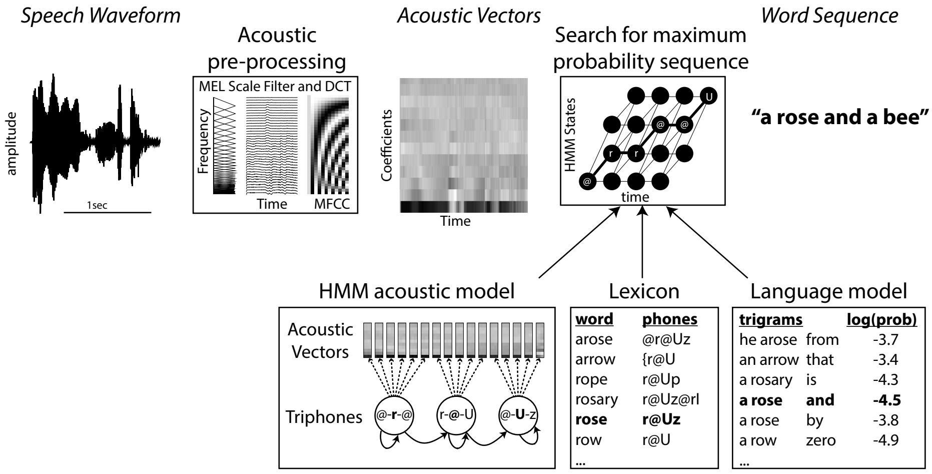
When faced with speech that is heard in a noisy background, is spoken in an unfamiliar accent or that contains out-of-vocabulary words, automatic systems struggle even more (for a humorous illustration of Apple's Siri system struggling to respond to a Scottish English speaker, see http://bit.ly/zY3eV9). Reviews of noise-robust machine speech recognition systems document the substantial engineering effort that has been dedicated to improving the performance of these systems (Li et al., 2014). Yet, even state-of-the-art performance still falls short of human listeners on all but the simplest of listening tasks. A recent series of engineering 'challenges' for noise-robust speech recognition have shown that for small vocabulary, closed-set tasks (reporting letters and/or digits), automated systems can approach human performance (see Barker et al., 2012). Accuracy remains high (>90%) even if the speech is quieter than the masking noise (a negative Signal to Noise Ratio, SNR). However, for a second challenge which involved a 5000 word vocabulary (Vincent et al, 2013) all the systems tested produced substantial numbers of errors for speech that is 9 dB louder than background noise; this SNR (+9dB) is typically fully intelligible for healthy human listeners (Miller et al., 1951). The best system showed a 15% keyword error rate at +9 dB SNR that increased to nearly 30% (Vincent et al, 2013) for speech masked by noise of equal amplitude (i.e. 0dB SNR). Healthy human listeners typically report connected speech with near perfect accuracy at 0dB SNR (Miller et al., 1951).

Despite these limitations current automatic systems are an impressive technological achievement and have reached a level of performance suitable for near-universal deployment in modern smartphones (with recognition typically achieved by processing speech in the "cloud" rather than in the phone itself). In the following section we will briefly describe the computational mechanisms by which these systems transcribe heard speech.

2. How machines recognize speech

The fundamental idea behind all successful machine speech recognition systems (following the pioneering work of Jelinek, 1976) is to treat the problem of spoken word recognition as a statistical problem – that of determining the most probable sequence of words given the acoustic speech input. This is exactly the same goal as explicitly stated in recent models of human speech perception (e.g. Norris & McQueen, 2008). In both cases, recognition is achieved by Bayesian perceptual inference with speech as the sensory input and the most likely word sequence as the desired output. However, ASR systems don't break this inference process into discrete parts (what are the most likely speech segments given the sounds heard, which are the most likely words given these speech segments, etc), in the same way that might be assumed for a bottom-up account of human speech recognition (see Mirman, this volume). Instead, typical ASR systems are designed in such a way as to combine acoustic and higher-level information throughout the recognition process using a single, integrated search process.

Figure 1 shows a schematic of a typical ASR system in which speech waveforms are first passed through an acoustic pre-processor to generate a sequence of vectors that represent the mel frequency spectrum (i.e. the energy profile over a frequency space similar to that of human hearing) of successive time windows of a spoken utterance. These acoustic features provide a relatively robust and compact description of the speech signal. These acoustic vectors provide the input to a search procedure (typically implemented using the Viterbi algorithm) that determines the most probable sequence of words that would have generated the observed sequence of acoustic vectors. In typical implementations, the search algorithm combines multiple sources of knowledge concerning (for instance) the probability of the sequence of acoustic vectors being generated by specific speech segments (acoustic model), which sequences of segments form real words (lexicon) and the relative likelihood of different word sequences (language model). Determining the sequence of words that is most likely to have generated an observed sequence of acoustic vectors allows an ASR system to report the most probably sequence of words contained in the speech waveform. The most probable word sequence can then be transcribed, used to drive a dialogue system, or for other purposes (see Young, 1996 for a more detailed review).



word	phones
arose	@r@Uz
arrow	{r@U
rope	r@Up
rosary	r@Uz@rl
rose	**r@Uz**
row	r@U
...	

trigrams		log(prob)
he arose	from	-3.7
an arrow	that	-3.4
a rosary	is	-4.3
a rose	**and**	**-4.5**
a rose	by	-3.8
a row	zero	-4.9
...		

Figure 1: Block diagram of a typical machine speech recognition system. Two key processes are: (1) Acoustic Pre-Processing in which the speech signal is passed through a MEL-scale filter bank followed by application of a Discrete Cosine Transformation to generate sequences of acoustic vectors (Mel Frequency Cepstral Coefficients) that represent the speech waveform, and (2) A search algorithm combines information from Hidden Markov Model (HMM) acoustic models, a lexicon and a language model to estimate the probability of different word sequences having generated the observed sequence of acoustic vectors. The maximum probability sequence is then returned as the systems best assessment of the sequences of words heard.

Most machine speech recognition systems achieve recognition by using several different representations for different sources of language knowledge (acoustic models, lexicon, language model); each of these components is typically chosen individually for their computational convenience and performance. An ASR system can be constructed from multiple different components because all these parts represent information in terms of probabilities that can be readily combined into a single search process.

One commonly used technique for relating the acoustic signal to speech segments are Hidden Markov Models (HMMs) which model the expected variation in the signal statistically and typically represent a single phoneme or a three-phoneme sequence (triphone). These HMMs provide a useful mechanism for dealing with sequences of acoustic vectors of variable lengths (e.g. due to differences in speech rate). Words are defined in the lexicon as sequences of acoustic models. Yet another, different knowledge source is used in computing the probability of word sequences; most systems use a language model that supplies the frequency of occurrence of words and the likelihood of different two or three word sequences (bigrams and trigrams) using a count of the conditional probability of successive words in a large sample of text or speech. Thus, despite agreement among engineers that Bayesian perceptual inference is the key to effective machine speech recognition, a number of different components can be used to implement this process (see Scharenborg et al., 2005 for a review from the perspective of human speech recognition).

During recognition, automatic systems typically use all these different sources of information (from the acoustic model, lexicon, and language model) at the same time. For any given utterance, the likelihood of the hypothesised word sequences (paths) is computed (e.g., using a graphical structure) that represents all the possible speech segments present in (some or all of) that utterance. Computing the likelihood of different paths involves multiplying the probabilities of sequences of segments so as to determine the probability of different words and word sequences. Thus acoustic, phonemic and word-based uncertainty is combined into a single, integrated optimization process. The length of the word sequence that will be recognised in a single optimisation is closely related to the complexity of the language model (for instance, if a trigram language model is used, then word sequences that typically contain 3 words will be optimised in a single search. This delay allows automatic systems to flexibly adjust the hypothesised word sequences, ultimately selecting the word sequence with the best match (i.e. highest probability) to have generated the input speech signal. This approach also means that unlike what is often assumed about human speech perception, automatic systems do not explicitly recognise speech sounds – that is, they do not have any knowledge of which speech segments were heard, only that specific words (that plausibly contain specific segments) were heard.

Although this approach yields effective systems for large-vocabulary continuous, speech recognition, it is often acknowledged in the automatic speech recognition community that the improvements in automatic speech recognition performance over the past decades can largely be credited to an increase in computing power and the availability of increasing amounts of suitable, high-quality speech material for training automatic speech recognition systems (e.g., Bourlard et al., 1996; Moore & Cutler, 2001),which both directly lead to more accurate acoustic models (De Wachter et al., 2007). For some time, however, increases in performance have slowed, reaching asymptote at a level that (as described in section 1) falls short of human performance. In order to break through this barrier, simply adding more training material will not help, nor is it to be expected that adding 'better' training material will help (see e.g., Kirchhoff & Schimmel, 2005, who used infant-directed speech to train automatic speech recognition systems). Instead, fundamentally new methodologies are needed (Bourlard et al., 1996; Moore, 2003; De Wachter et al., 2007). In this chapter we will discuss three recent developments in ASR which are (to varying degrees) inspired by human speech recognition and that might contribute to future progress in ASR.

3. Alternative representations of speech signals

In implementing automatic speech recognition systems, certain practical decisions have to be made concerning the representations used at different levels of the system. For example, as described above speech waveforms recorded by a microphone are transformed into sequences of acoustic vectors which are used to train HMM-based acoustic models. The representations that are typically used are Mel-Frequency Cepstral Coefficients (MFCCs, Davis & Mermelstein, 1980), based on a non-linear spectral representation for each time window of the short-term Fourier spectrum of speech (i.e. the spectrum of a spectrum, see Figure 1). These are often described as inspired by certain characteristics of the human auditory system (the Mel-frequency scale is based on the frequency spacing of human auditory filters at different centre frequencies). This proves to be an effective representation of the speech signal for HMM-based systems since it removes a great deal of redundant information from the speech stream and excludes information that is irrelevant for the recognition of words, such as pitch and continuous background noise. This form of data reduction is effective for HMM acoustic models working with clear speech since HMMs have only a limited ability to deal with non-linearity or redundancy. More detailed (but also more redundant) acoustic representations such as the output of an auditory filterbank can be used. For example, it has also been proposed that more robust recognition performance in difficult listening conditions might be achieved with other representations of the speech waveform (Li et al., 2014).

Another form of representation that is commonly incorporated into automatic recognition systems is the phoneme. Phonemes are commonly used as the mediating step between acoustic signals and specific words, i.e. the acoustic models represent phonemes or triphones and words are made from sequences of these units. This approach implements what is known as the 'beads on a string' model of speech perception (Ostendorf, 1999). Although this model works satisfactorily for carefully produced speech, it runs into problems with more naturalistic speech. This is mainly due to the high pronunciation variability in naturalistic speech (e.g. due to coarticulation and phonetic reduction processes “stand back” can be pronounced /stam bak/ in connected speech, see Gaskell & Marslen-Wilson, 1996). The strict, segmental nature of phoneme-based acoustic models limits their sensitivity to the fine-grained acoustic detail of speech. For example, in deciding whether the words “grade A” or “grey day” is a better transcription of the sounds /greidei/ these systems overlook subtle acoustic cues (such as syllable duration, stress patterns, coarticulation, allophonic variation, etc) that provide phonetic evidence to distinguish between sequences of sounds that occur within a single word or straddle word boundaries. Such cues are distributed over time, and do not easily feature in phoneme based HMM models but have nonetheless been shown to be used by human listeners (Davis et al., 2002; Salveda et al, 2003; Shatzman & McQueen, 2006a/b; see Scharenborg, 2010 for a review).

To overcome these problems alternative representations have been proposed, such as representations based on articulatory features (King & Taylor, 2000; Kirchhoff, 1999). These alternative accounts are often motivated with respect to the characteristics of the human speech recognition system in which feature representations are often proposed to mediate between acoustic and lexical representations of speech (e.g. Jakobson, Fant, & Halle, 1952; see also Lahiri & Reetz, 2010; Marslen-Wilson & Warren, 1994; Johnsrude & Buchsbaum, this volume). Articulatory or articulatory-acoustic features (hereafter, AFs) describe properties of articulatory events – that is the lip, mouth and tongue movements that speakers make when producing speech sounds. However, these are typically not embodied in detailed mechanical descriptions, but rather abstract classes that characterise the most essential aspects of the articulatory properties of speech sounds such as manner and place of articulation, tongue position, and voicing. With this type of feature, speech can be represented without necessarily assuming a sequence of discrete segments. Consequently, fine-phonetic detail such as nasalisation of a vowel preceding a nasal sound (such as in the vowel of the word "ban") can contribute to identification of nasal segments (/n/), whilst not creating difficulties for the identification of the vowel /a/ (see Lahiri & Marslen-Wilson, 1991; Hawkins, 2003).

Many different approaches have been investigated for incorporating AFs into automatic speech recognition systems, though to-date none of these have been incorporated into commercial ASR systems. These include using artificial neural networks (King & Taylor, 2000; Kirchhoff, 1999; Wester, 2003), HMMs (Kirchhoff, 1999), linear dynamic models (Frankel, 2003), dynamic Bayesian networks (Livescu et al., 2003), and support vector machines (Scharenborg, Wan, & Moore, 2007) to replace HMM-based acoustic models. AF classifiers have been used to improve speech recognition performance in adverse conditions (Kirchhoff, Fink, & Sagerer, 2002; Kirchhoff, 1998), to build language independent phone recognizers (Siniscalchi & Lee, 2014), and to improve computational models of human word recognition (Scharenborg, 2010). This last model is particularly helpful for illustrating how the use of articulatory features and duration representations can simulate human data on the recognition of words in which lexical segmentation creates ambiguity (as for onset-embedded words like *cap* in *captain*, cf. Davis et al., 2002; Salverda et al., 2003; or segmentation minimal pairs like “grade A” and “grey day”, cf. Nakatani & Dukes, 1977; Shatzman & McQueen, 2006a/b).

Although promising, the lack of large training corpora that label the speech signal in terms of AF values hampers the further development of AF-based systems (only one small training set is available, created during the 2004 Johns Hopkins Summer Workshop, Livescu et al., 2007). The most popular corpus for research into AF classification is the standard TIMIT database designed for comparison of more conventional ASR systems (Garofolo, 1988). This is a corpus of read American English consisting of high quality manually created phonetic transcriptions using a large set of phonetic labels. Consequently, training and testing of AF classifiers is generally achieved by starting from data that is labelled at the phoneme level and replacing phoneme labels with their (canonical) AF values. These AF values change synchronously at the phoneme boundaries, losing a large part of the potential for AF representations as an alternative to segmental representation (Schuppler et al., 2009).

An alternative or complementary proposal to using sub-phonemic, articulatory features in ASR is that articulatory features are combined into larger syllabic units during recognition (see, for instance, Greenberg, 1999 for a prominent example). It has been proposed, for instance, that many forms of pronunciation variability (such as duration changes) can be more effectively modelled using syllables rather than phonemes as the unit of representation (Greenberg et al, 2003). However, to date, there are few viable ASR systems that have been built in this way (see Kirchoff, 1996; Puurula & van Compernelle, 2010, for attempts). One problem with this approach is that typical ASR systems use segment level transcriptions to link acoustic models to a lexicon of known words. A syllable based model that eschews segmental representations would have no way to identify the syllables in low-frequency monosyllabic words other than by learning from the exemplars of these words that occur in the training set. In contrast, a system that works with segmental or AF representations can recognise low-frequency words as sequences of more frequent segments and is therefore likely to be more successful at word recognition.

These debates in the ASR literature concerning the units of representation that are most effective for speech recognition parallel long-standing debates in the literature on human speech recognition concerning the nature of speech representations (see Johnsrude and Buchsbaum, this volume, or Goldinger & Azuma, 2003). An alternative approach, however, would be to allow the automatic recognition system to determine which unit or units of representation most reliably mediate between the acoustic signal and word recognition. One way to achieve this is to have automatic systems break the input sequences of acoustic features into either pre-defined 'units' or 'units' that are automatically learned and can then be used in the recognition of words (e.g., Aimetti et al., 2009; De Wachter et al., 2007). An alternative approach is to use neural network learning algorithms to 'discover' suitable intermediate representations between speech and words. The next section of this chapter will review historical and more recent approaches to ASR using neural networks. However, the majority of existing neural network based ASR systems do not use neural networks to achieve word recognition directly from acoustic representations but rather use neural networks to replace the HMM-based acoustic models in existing systems. Thus, the flow diagram depicted in Figure 1, with minor modifications, remains an accurate description of most current ASR systems.

4. Neural networks for machine speech recognition

Neural networks are multi-layer systems of simple processing units that compute the weighted sum of their inputs, which is then passed through a non-linear function and output to subsequent levels of processing (see Bishop, 1996; for an introduction). These simple, neurally inspired systems have a long history. Their adoption and capacities for tasks such as automatic speech recognition have been largely due to the development and refinement of learning algorithms that are capable of setting the weights on the connections between units so as to solve specific problems (e.g. mapping from sequences of acoustic vectors to phonemes or words). Among the earliest of these was the perceptron learning procedure of Rosenblatt (1958). However, more significant advances followed the development (or rediscovery) of learning by back-propagation of error by Rumelhart, Hinton and Williams (1986) along with subsequent extensions of this procedure to train recurrent neural networks – that is systems with internal states that retain a 'history' of past input and that can therefore process signals (such as speech) that unfold over time (Pearlmutter, 1995). Critically, back-propagation and other, similar learning algorithms can be used to train networks with hidden units. These allow for the input to be transformed into mediating representations so that these networks can learn non-linearly separable mappings that evade simpler methods such as perceptrons or HMMs (see Bishop, 1996 for a detailed presentation of linear separability).

As part of a flurry of interest in neural networks that followed the two volume "PDP books" in the 1980s (McClelland & Rumelhart, 1986; Rumelhart & McClelland, 1986) many researchers explored the possibility of using either static or recurrent neural networks in machine speech recognition (see Lippman, 1989 for an early review of these efforts). However, these systems often failed to achieve sufficient scale (in terms of the size of training materials) or accuracy (e.g. phoneme identification performance) to supplant existing HMM-based systems. A few notable successes were achieved by using hybrid systems in which recurrent neural networks with a single hidden layer were used to compute phoneme probabilities from speech signals which could then be interfaced with conventional Viterbi-based search procedures (see, for instance, Robinson, 1994; Bourlard & Morgan, 1994). Despite these early demonstrations, however, the neural network components of these hybrid systems were hard to train due to the limited speed of workstation computers at that time. Furthermore, the performance advantages offered by systems with a single hidden layer were not sufficient to render more conventional HMM systems obsolete and training procedures for multi-layer systems were of limited ability. Hence, these recurrent network systems were in time replaced by more tractable HMM-based systems as described in section 2 of the present chapter.

Recent years have seen a resurgence of interest in the use of neural networks for machine speech recognition. In part this is the result of increases in computer processing speed, particularly by using graphics processors to perform fast vector computations during training. One striking and influential demonstration by Mohamed, Dahl & Hinton (2009) showed that a deep neural network (DNN, i.e. a neural network with multiple layers of hidden units between the input and output) could substantially improve on state-of-the-art HMM scores on the TIMIT phoneme identification task. This success has led many groups to make updated versions of the hybrid systems that were used in the 1990s by using neural networks to replace the traditional HMM-based acoustic models (see Hinton et al., 2012 for a review). These DNN-based systems are trained using large sets of phonetically labelled speech signals to output the posterior probability of different HMM states (phonemes or triphones, for example) given a sequence of acoustic states as input. We will therefore consider the critical elements of this advance and consider the parallels between this approach and proposals made for human speech recognition – a domain in which neural network or connectionist approaches remained popular throughout the intervening decades (as exemplified by simple recurrent network models, see Gaskell & Marslen-Wilson, 1997; Mirman this volume).

The modern resurgence of interest in DNN systems for phoneme identification arises not only from the increased speed of modern computers, but also from the development of new and more robust methods for training DNNs. One reason for the success of DNNs for classification tasks (such as phoneme recognition) is the use of generative, pre-training schemes in which a DNN learns (in an unsupervised fashion) to represent the acoustic characteristics of the input. Input to these models is often supplied using the same Mel-Frequency Cepstral Coefficient (MFCC) representations used in HMM-based systems, though other, more redundant auditory representations (such as the output of auditory filter banks) have been tried with some success (Hinton et al, 2012). The training procedure used is hierarchical – a single layer is trained to represent first-order dependencies in the acoustic vectors, before an additional layer is added to represent dependencies among these first order dependencies, and then a third layer and so on. Critically, these models are generative – connectivity is bidirectional and typical training algorithms (e.g. the contrastive divergence algorithm, Hinton et al, 2006) alternate between "wake" phases in which the model derives internal representations from the input, and "sleep" phases in which the model uses those internal representations to reconstruct input sequences similar to those that were presented (see Hinton, 2014 for a overview). These procedures provide an effective procedure for discovering compact representations of sequences of acoustic feature vectors.

For machine speech recognition, however, it is not sufficient to derive a robust, and compact representation of the speech signal. These representations also have to be categorized into discrete units (such as single phonemes, N-phones, features, syllables, etc), in order to make contact with higher-level representations such as words. Hence, the hierarchical stack of units and connections that were trained to represent and reconstruct speech signals in a DNN are interfaced with a final layer of units with an output function suitable for classifying speech signals into unique categories (typically triphones). The entire stack of units (including the lower-level stages that were originally trained in an unsupervised manner) are then subjected to discriminative or supervised training using the back-propagation learning algorithm (Rumelhart et al., 1986). The full system is then able to output the probability of different units in the acoustic signal (typically N-phone probabilities) with an accuracy unmatched by HMM-based systems (see Hinton et al., 2012). The key advance provided by DNNs, relative to both HMMs and recurrent networks with a single layer of hidden units is that these networks provide a powerful mechanism for learning multiple layers of non-linear features (see Hinton, 2014 for discussion). This success has led to the adoption of DNN methods by many of the major commercial speech recognition systems (see McMillan, 2013 for an accessible introduction).

From the perspective of computational modelling of human speech recognition, these two stages of training an acoustic model (generative pre-training and discriminative training) are reminiscent (in their goal, if not in their methods), of connectionist approaches to human spoken language acquisition (see Mirman, this volume). Building on recurrent network simulations reported in Elman (1990), a number of authors have proposed that early stages of speech acquisition (during the first year of life) are well-explained by training recurrent neural networks to predict subsequent segments in extended sequences of speech sounds (Cairns et al., 1997; Christiansen, Allen & Seidenberg, 1998). These self-supervised neural networks develop internal representations that capture important forms of linguistic structure such as words in artificially coded speech sequences, and periods of accurate and inaccurate prediction reflect knowledge of likely words in connected speech signals. The second stage of supervised learning used in these DNN systems is also reminiscent of procedures used in training connectionist or neural network models of spoken word recognition (such as the Distributed Cohort Model, Gaskell & Marslen-Wilson, 1997, or other similar models using localist representations of spoken words, Norris, 1990; Davis, 2003, see Mirman chapter). Interestingly, these recurrent network systems appear to perform better if both forms of learning (unsupervised prediction, and supervised lexical identification) are combined in a single system (Davis, 2003; Mirman et al., 2010).

Despite the gratifying success of neurally-inspired components in machine speech recognition systems many of these systems still make unrealistic assumptions about how the temporal structure of the speech signal should be coded. The DNNs described so far, mostly use separate sets of input units to code a sequence of acoustic vectors. That is, they use different units and connections information that occurs at the present and previous time points; they also retain a veridical (acoustic) representation of the preceding acoustic context. Thus, these models use an unanalysed acoustic context for the recognition of the most likely speech segment in the current acoustic vector (as in Time-Delay Neural Networks described by Waibel et al, 1989). This is a spatial method of coding temporal structure (similar to that used in the TRACE model, McClelland & Elman, 1986). Spatial coding seems unrealistic as a model of how temporal structure and acoustic context is processed during speech perception; Humans don't use different auditory nerve fibres or cortical neurons to process sounds that are presented at different points in time, but rather the same neurons provide input at all points in time, and perception is supported by internal representations that retain relevant context information.

A more appropriate method for coding temporal structure therefore involves using recurrent neural networks, in which input is presented sequentially (one acoustic vector at a time) and activation states at the hidden units provide the temporal context required to identify the current input which can be trained with variants of back-propagation (see Elman, 1990; Pearlmutter, 1995). Recurrent neural networks were initially used successfully in phoneme probability estimation (e.g. Robinson, 1994), but were found to be difficult to train, particularly when long-distance dependencies must be processed in order to identify speech signals (for instance, if input from several previous time steps must be used to inform the current input). Sequences in which there are long delays from when critical information appears in the input and when target representations permit back-propagation of error, require that weight updates be passed through multiple layers of units (one for each intervening time-step) during training. These additional intervening units make it more likely that error signals will become unstable (since error gradients can grow exponentially large, or becoming vanishingly small, see Hochreiter et al, 2001). Various solutions to this problem of learning long-distance temporal dependencies have been proposed including schemes for incremental learning of progressively longer-distance dependencies (e.g. Elman, 1993). Perhaps the most powerful solution, however, comes from "Long Short Term Memory" networks proposed by Hochreiter & Schmidhuber (1997) in which error signals are preserved over multiple time points within gated memory circuits. These systems achieve efficient learning of long-distance dependencies and are now being used in deep neural network systems for acoustic modelling (see Beaufais, 2015).

Despite the successful deployment of these neural networks, their incorporation into existing ASR systems has still largely come from replacing single components of existing systems with DNNs and not from an end-to-end redesign of the recognition process. For example, DNN have been used to replace the HMM acoustic model shown in Figure 1. However, this still requires that the phoneme classification output of a neural network is transformed into standard HMM states (corresponding to phonemes), and a search algorithm is used to combine these HMM states into word sequences constrained by an N-gram based language models (essentially the same hybrid connectionist approach proposed in Bourlard & Morgan, 1994). More recently, some authors have begun to explore the possibility of end-to-end neural network based speech recognition systems (e.g. Graves & Jaistly, 2014; Chorowski et al., 2014). These systems have not so far been sufficiently successful (or computationally tractable) to operate without a traditional N-gram-based language model. Furthermore, while DNN-based language models have been proposed in other contexts (e.g. for machine translation systems, Cho et al., 2014) these have rarely been interfaced to a perceptual system based around a DNN. We note, however, that end-to-end computational models of human word recognition have been constructed using a recurrent neural network (e.g. Gaskell & Marslen-Wilson, 1997). This 'distributed cohort' model uses back-propagation to map from (artificially coded) speech segments to meaning. While this model is small in scale, and unable to work with real speech input recent progress in the use of neural networks for ASR suggest that this model could be developed further.

5. Perceptual learning in human and machine speech recognition

Perhaps the most significant challenge for machine speech recognition is that the identity of speech sounds is not only determined by the acoustic signal, but also by the surrounding context (acoustic, lexical, semantic, etc) in which those sounds occur and by knowledge of the person who produced these sounds (their vocal tract physiology, accent etc). The optimal use of contextual information in recognition is not easily achieved by using either an HMM or a time-delay DNN for acoustic modelling in ASR systems. In both cases, only a relatively short period of prior acoustic context is encoded in the input to the acoustic models, and perceptual hypotheses for the identity of the current segment are determined (bottom-up) only on the basis of this acoustic input. For this reason, ASR systems defer decisions concerning the identity of specific speech segments until these sub-lexical perceptual hypotheses can be combined with higher-level information (such as knowledge of likely words, or word sequences). As shown in Figure 1, identification of speech sounds in ASR systems arises through the combination of acoustic models with a lexicon and language model so that lexical and semantic/syntactic context can be used to support speech identification.

Human recognition shows similar lexical and sentential influences on segment identification. This has been shown by changes to phoneme categorization boundaries that favour real words or meaningful sentences. For example, a sound that is ambiguous between a /t/ and /d/ will be heard differently in syllables like “task” or “dark” (since listeners disfavour nonword interpretations like “dask” or “tark”, i.e. the Ganong effect; Ganong, 1980). Furthermore, even when disambiguating information is delayed beyond the current syllable (as for an ambiguous /p/ and /b/ at the onset of “barricade” and “parakeet”) listeners continue to use lexical information to resolve segmental ambiguities in a graded fashion (McMurray et al., 2009). Sentence level meaning that constrains word interpretation has also been shown to modify segment perception (Borsky et al., 1998). Thus, human listeners, like machine recognition systems delay phonemic commitments until higher-order knowledge, including lexical and semantic information can be used to disambiguate.

However, unlike human listeners, typical ASR systems do not change their subsequent identification of speech segments as a consequence of lexically- or semantically-determined disambiguation. As first shown by Norris, McQueen & Cutler (2003, see Samuel & Kraljic, 2009 for a review) a process of perceptual learning allows human listeners to use lexical information to update or retune sub-lexical phoneme perception. That is, hearing an ambiguous /s/-/f/ segment at the end of a word like “peace” or “beef” that constrains interpretation leads to subsequent changes in the perception of an /s/ or /f/ segment heard in isolation. Human listeners, infer that they are listening to someone that produces specific fricatives in an ambiguous fashion and change their interpretations of these sounds accordingly (see Kraljic & Samuel, 2006; Eisner & McQueen, 2006 for contrasting findings, however, concerning generalization between speakers).

Recent evidence suggests that for human listeners, perceptual learning only arises for ambiguous segments that occur towards the end of a word (Jesse & McQueen, 2011). Perceptual learning is thus absent for word-initial /s/-/f/ ambiguities even in strongly constraining contexts like “syrup” or “phantom” despite successful identification of the spoken words in these cases. Although human listeners can delay making commitments to specific phonemes in order to correctly identify words, they appear not to use these delayed disambiguations to drive perceptual learning. These observations suggest mechanisms for perceptual learning that are driven by prior knowledge of upcoming segments and not solely by word identification.

In combination, then, these learning effects point to a form of perceptual flexibility that is often critical for successful human speech recognition. Listeners are adept at using information gained from previous utterances to guide processing of future utterances. In real world listening situations, this learning process is most apparent when listeners hear strongly accented speech. Accented speech may contain multiple segments for which the form of perceptual learning described previously is required. Laboratory studies have shown rapid gains in the speed and accuracy of word identification following relatively short periods of exposure to accented speech (Clarke & Garrett, 2004; Adank & Janse, 2010).

One way of describing this process is as a form of (self-) supervised learning similar to that used in training deep neural networks (see Norris et al., 2003; Davis et al, 2005). For human listeners, lexical identification provides knowledge of the segments that were presented in the current word. This knowledge is then used in a top-down fashion to modify the mapping from acoustic representations to segment identity such that a previously ambiguous acoustic input is more easily identified in future. While this process is similar to the supervised learning algorithms used in training DNNs, the neural networks in current ASR systems do not use such mechanisms during recognition. The procedures that are used to train the weighted connections in these systems require batched presentation of large quantities of training data including (for discriminative training) external signals that supply frame-by-frame ground-truth labels of the phonemic content of speech signals. When these systems are used to recognise speech they operate with these learning mechanisms disabled (that is, the weighted connections between units remain the same irrespective of the utterance that is being recognised).

One obstacle to including perceptual learning mechanisms in ASR systems is therefore that ASR systems would need to derive top-down supervisory signals without external guidance. That is, the system must not only recognise words, but also determine whether or not recognition is sufficiently accurate to support changes to the mapping from acoustic vectors to segments (since it's better not to learn from incorrect responses). This introduces a further requirement; specifically that the system has an internally-derived measure of confidence in its own recognition. At present, however, measures of confidence have not been used for this purpose (see Jiang, 2005 for a review of attempts to derive confidence measures from existing ASR systems). There is, however, some experimental evidence that recognition confidence may modulate the efficacy of human perceptual learning (see Drozdova et al., 2015; Zhang & Samuel, 2014).

Mechanisms for adaptation to speaker-specific characteristics have however been incorporated into HMM-based machine recognition systems. These typically operate by including additional hyper parameters that are associated with specific utterances or speakers heard during training (Woodland, 2001; Yu & Gales, 2007). Techniques such as Maximum a Posteriori (MAP) parameter estimation and Maximum Likelihood Linear Regression (MLLR) can then be used adapt the trained model parameters or to establish hyper-parameters that optimize perception of utterances from a new speaker. These methods permit adaptation to a new speaker based on a more limited number of utterances than would otherwise be required. Similar maximum likelihood methods have also been used in accommodating speakers with different length vocal tracts (which systematically change formant frequencies). However, a more straight-forward frequency warping can also be used to adapt to novel speakers (Lee & Rose, 1998).

One distinction between machine and human adaptation that we wish to draw, however, is between machine recognition systems that adapt by using relevant past experience of similar speakers and human listeners that show rapid learning even when faced with entirely unfamiliar (fictitious) accents. For instance, in studies by

Adank & Janse (2010) young listeners showed substantial improvements in their ability to comprehend a novel accent created by multiple substitutions of the vowels in Dutch (e.g. swapping tense and lax vowels, monopthongs and dipthongs, etc). Improvements in comprehension were even more rapid when listeners were instructed to imitate the accented sentences (Adank et al., 2010). These behavioural experiments point to a form of adaptation that can operate even when listeners have no relevant past experience of any similar accent. That this is a form of supervised learning is also apparent from research showing that accent adaptation is more rapid for listeners that receive supervisory information from concurrent written subtitles (Mitterer & McQueen, 2009).

Human listeners also show perceptual learning when faced with extreme, or unnatural forms of degraded speech. For example, perceptual learning occurs when listeners hear speech that has been artificially time-compressed to 35% of the original duration (Mehler et al., 1993), or noise-vocoded to provide just a handful of independent spectral channels (vocoded speech, Davis et al., 2005), or resynthesized using only three harmonically unrelated whistles (sine-wave speech, Remez et al., 2011). In all these cases, listeners rapidly adapt despite having had essentially no relevant prior exposure to other similar forms of speech. Once again, many of these forms of learning are enhanced by prior knowledge of speech content (e.g. written-subtitles, or clear speech presentations) that precede perception of degraded speech (e.g. Davis et al., 2005; Hervais-Adelman et al., 2008) further suggesting supervisory mechanisms involved in perceptual learning.

In sum, this evidence suggests that rapid and powerful learning processes contribute to successful identification of accented and degraded speech in human listeners. It remains to be seen whether incorporating a similar form of self-supervised learning would enhance the performance of machine recognition systems. In explaining the abilities of human listeners, computational models of spoken word recognition have already been proposed that can adjust their internal processes to simulate perceptual learning of ambiguous speech segments (HebbTRACE: Mirman, McClelland & Holt, 2006; Kleinschmidt & Jaeger, 2014). However, one interesting, and under-explored aspect of these models concerns the situations in which such rapid learning is possible. We have noted that accurate prior knowledge of the likely identity of upcoming speech segments is a necessary condition for perceptual learning to occur (cf. Jesse & McQueen, 2011; Davis et al., 2005). Predictive coding mechanisms may provide one proposal for how these findings can be accommodated in models of human speech recognition (Sohoglu et al., 2012; Gagnepain et al., 2012): accurate predictions for upcoming speech signals are reinforced to drive perceptual learning, whereas speech signals that lead to large prediction errors provide a novelty signal to drive encoding of unfamiliar words.

6. Summary

This chapter has described the inner-workings of machine speech recognition systems that have already transformed our day-to-day interactions with computers, smartphones, and similar devices. Improvements in the effectiveness and convenience of voice input seems set to continue; we imagine that our children will in time be amused at our generation's antiquated attachment to QWERTY keyboards. However, the ASR systems that we have described still fall short of human levels of recognition performance. Substantial improvements will be required if our communication with machines is to be as seamless as it is with our friends and family.

We have offered three distinct proposals for key aspects of human speech recognition that could inspire future developments in machine recognition systems. Specifically, we have proposed that it is worth exploring ASR systems that: (1) relax the assumption that speech is comprised of a sequence of discrete and invariant segments (phonemes), (2) operate in an end-to-end fashion using neural network components, and (3) are able to learn from their own successes at recognition. We hope that these changes might allow for further progress in achieving accurate and robust machine speech recognition. However, we also acknowledge that existing systems are already good enough for day-to-day use by millions of people around the world. There is much for researchers in human speech recognition to gain from understanding the computational mechanisms that have achieved these successes. We hope that the overview of the underlying technology in the present chapter allows psycholinguists to learn from the successes of engineers and computer scientists working to improve ASR systems.

References

Adank, P., Hagoort, P., & Bekkering, H. (2010). Imitation improves language comprehension. *Psychological Science*, *21*(12), 1903–9. doi:10.1177/0956797610389192

Adank, P., & Janse, E. (2010). Comprehension of a novel accent by young and older listeners. Psychology and Aging, 25, 736–740.

Aimetti, G., ten Bosch, L., & Moore, R. K. (2009). Modelling early language acquisition with a dynamic systems perspective, 9th Int. Conf. on Epigenetic Robotics. Venice.

Barker, J., Vincent, E., Ma, N., Christensen, H., Green, P., Barker, J., … Pascal, T. (2013). The PASCAL CHiME Speech Separation and Recognition Challenge. *Computer Speech & Language*, *27*(3), 621–633.

Beaufais, F. (2015). The neural networks behind Google Voice transcription. Downloaded from http://googleresearch.blogspot.co.uk/2015/08/the-neural-networks-behind-google-voice.html

Bourlard, H. A., & Morgan, N. (1994). Connectionist speech recognition: A hybrid approach. Boston: Kluwer.

Bourlard, H., Hermansky, H., Morgan, N., 1996. Towards increasing speech recognition error rates. Speech Communication 18, 205–231.

Borsky, S., Tuller, B., & Shapiro, L. (1998). "How to milk a coat:" The effects of semantic and acoustic information on phoneme categorization. Journal of the Acoustical Society of America, 103, 2670–2676.

Cairns, P., Shillcock, R., Chater, N., & Levy, J. (1997). Bootstrapping Word Boundaries: A Bottom-up Corpus-Based Approach to Speech Segmentation. Cognitive Psychology, 33(2), 111–53. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/9245468

Carey, M.J., Quang, T.P., 2005. A speech similarity distance weighting for robust recognition. Proceedings of Interspeech, Lisbon, Portugal, pp. 1257-1260.

Cho, K., van Merrienboer B., Caglar Gulcehre, Fethi Bougares Holger Schwenk Yoshua Bengio, (2014), Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation. Conference on Empirical Methods in Natural Language Processing (EMNLP 2014)

Jan Chorowski, Dzmitry Bahdanau, Kyunghyun Cho, & Yoshua Bengio. (2014). End-to-end Continuous Speech Recognition using Attention-based Recurrent NN: First Results, 1–10. Retrieved from http://arxiv.org/abs/1412.1602v1

Christiansen, M. H., Allen, J., & Seidenberg, M. S. (1998). Learning to Segment Speech Using Multiple Cues: A Connectionist Model. *Language and Cognitive Processes*, *13*(2-3), 221–268. doi:10.1080/016909698386528

Clarke, C. M., & Garrett, M. F. (2004). Rapid adaptation to foreign-accented English. *The Journal of the Acoustical Society of America*, *116*(6), 3647–3658. doi:10.1121/1.1815131

Cooke, M., 2006. A glimpsing model of speech recognition in noise. Journal of the Acoustical Society of America 119 (3), 1562–1573.

Cutler, A., Robinson, T., 1992. Response time as a metric for comparison of speech recognition by humans and machines. Proceedings of ICSLP, Banff, Canada, pp. 189-192.

Davis, M. H. (2003) "Connectionist modelling of lexical segmentation and vocabulary acquisition" in Quinlan, P. (Ed) Connectionist models of development: Developmental processes in real and artificial neural networks. Psychology Press, Hove, UK.

Davis, M. H., Johnsrude, I. S., Hervais-Adelman, A., Taylor, K., & McGettigan, C. (2005). Lexical information drives perceptual learning of distorted speech: evidence from the comprehension of noise-vocoded sentences. *Journal of Experimental Psychology. General*, *134*(2), 222–41. doi:10.1037/0096-3445.134.2.222

Davis, M. H., Marslen-Wilson, W. D., & Gaskell, M. G. (2002). Leading up the lexical garden path: Segmentation and ambiguity in spoken word recognition. *Journal of Experimental Psychology: Human Perception and Performance*, *28*(1), 218–244. doi:10.1037//0096-1523.28.1.218

Davis, S., & Mermelstein, P. (1980). Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, *28*(4). doi:10.1109/TASSP.1980.1163420

De Wachter M., Matton M., Demuynck K., Wambacq P., Cools R., and Van Compernolle, D. Template Based Continuous Speech Recognition. IEEE Transactions on Audio, Speech and Language Processing, volume 15, pages 1377-1390, May 2007.

Drozdova, P., van Hout, R., Scharenborg, O. (2015). The effect of non-nativeness and background noise on lexical retuning. Proceedings of the International Congress of Phonetic Sciences, Glasgow, UK.

Eisner, F., & McQueen, J. M. (2006). Perceptual learning in speech: Stability over time. *The Journal of the Acoustical Society of America*, *119*(4), 1950. doi:10.1121/1.2178721

Elman, J. (1990). Finding structure in time. *Cognitive Science*, *14*(2), 179–211. doi:10.1016/0364-0213(90)90002-E

Elman, J. L. (1993). Learning and development in neural networks: the importance of starting small. *Cognition, 48*, 71-99.

Frankel, J., Wester, M., King, S. (2007) Articulatory feature recognition using dynamic Bayesian networks. Computer Speech and Language, 21 (4), pp. 620-640

Gagnepain, P., Henson, R. N., & Davis, M. H. (2012). Temporal predictive codes for spoken words in auditory cortex. *Current Biology : CB*, *22*(7), 615–21. doi:10.1016/j.cub.2012.02.015

Ganong, W. F. (1980). Phonetic categorization in auditory word perception. *Journal of Experimental Psychology: Human Perception and Performance, 6*, 110- 125.

Garofolo, J. S. 1988. Getting started with the DARPA TIMIT CD-ROM: An acoustic phonetic continuous speech database. Technical report, National Institute of Standards and Technology (NIST).

Gaskell, M. G., and Marslen-Wilson, W.D. (1996) Phonological variation and inference in lexical access. Journal of Experimental Psychology: Human Perception and Performance , 22, 144-158.

Gaskell, M. G., & Marslen-Wilson, W. D. (1997). Integrating Form and Meaning: A Distributed Model of Speech Perception. *Language and Cognitive Processes*, *12*(5-6), 613–656. doi:10.1080/016909697386646

Goldinger, S. D., & Azuma, T. (2003). Puzzle-solving science: The quixotic quest for units in speech perception. *Journal of Phonetics*, *31*(3-4), 305–320. doi:10.1016/S0095-4470(03)00030-5

Graves, A., & Jaitly, N. (2014). Towards End-To-End Speech Recognition with Recurrent Neural Networks. *JMLR Workshop and Conference Proceedings*, *32*(1), 1764–1772. Retrieved from http://jmlr.org/proceedings/papers/v32/graves14.pdf

Greenberg, S. (1999), Speaking in shorthand: a syllable-centric perspective for understanding pronunciation variation. Speech Commun. 29: 159–176 (1999).

Greenberg, S., Carvey, H., Hitchcock, L., and Chang, S. (2003). Temporal properties of spontaneous speech – a syllable-centric perspective. *J. Phon.* 31, 465–485. d

Hain, T., Burget, L., Dines, J. , Garau, G., Karafiat, M., Lincoln, M., Vepa J., and Wan, V. (2007) The AMI system for the Transcription of Speech in Meetings,in: *Proceedings of the 32nd International Conference on Acoustics, Speech, and Signal Processing* (ICASSP 2007).

Hawkins, S. (2003). Roles and representations of systematic fine phonetic detail in speech understanding. *Journal of Phonetics*, *31*(3-4), 373–405. doi:10.1016/j.wocn.2003.09.006

Hervais-Adelman, A., Davis, M. H., Johnsrude, I. S., & Carlyon, R. P. (2008). Perceptual learning of noise vocoded words: effects of feedback and lexicality. *Journal of Experimental Psychology. Human Perception and Performance*, *34*(2), 460–74. doi:10.1037/0096-1523.34.2.460

Hilger, F., Ney, H. (2006). Quantile Based Histogram Equalization for Noise Robust Large Vocabulary Speech Recognition. IEEE Transactions on Audio, Speech, and Language Processing, Vol. 14, No. 3.

Hinton, G., Deng, L., Yu, D., Dahl, G., Mohamed, A.-R., Jaitly, N., … Kingbury, B. (2012). Deep Neural Networks for Acoustic Modeling in Speech Recognition. *Signal Processing Magazine, IEEE*, *29*(november), 82–97. doi:10.1109/MSP.2012.2205597

Hinton, G. E., Osindero, S. and Teh, Y. (2006) A fast learning algorithm for deep belief nets. Neural Computation 18, pp 1527-1554

Hinton, G. E. (2014) Where do features come from?.Cognitive Science, Vol. 38(6), pp 1078-1101.

Sepp Hochreiter, Yoshua Bengio, Paolo Frasconi, Juergen Schmidhuber. (2001) Gradient flow in recurrent nets: the difficulty of learning long-term dependencies. In S. C. Kremer and J. F. Kolen, eds., A Field Guide to Dynamical Recurrent Neural Networks. IEEE press.

Hochreiter, S., Hochreiter, S., Schmidhuber, J., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*(8), 1735–80. doi:10.1162/neco.1997.9.8.1735

Jakobson, R., Fant, G. M. C. & Halle, M. (1952). *Preliminaries to Speech Analysis: The Distinctive Features and their Correlates*, MIT Press, Cambridge, MA, U.S.A.

Jelinek, F. (1976). Continuous speech recognition by statistical methods. Proceedings of the IEEE, 64(4):532-536.

Jesse, A., & McQueen, J. M. (2011). Positional effects in the lexical retuning of speech perception. *Psychonomic Bulletin & Review*, *18*(5), 943–50. doi:10.3758/s13423-011-0129-2

Jiang, H. (2005). Confidence measures for speech recognition: A survey. *Speech Communication*, *45*(4), 455–470. doi:10.1016/j.specom.2004.12.004

Juneja, A. (2012). A comparison of automatic and human speech recognition in null grammar. *Journal of the Acoustical Society of America,* 131(3), *EL256-261.*

Jurafsky, Daniel; James H. Martin (2009). *Speech and language processing: an introduction to natural language processing, computational linguistics, and speech recognition*. Prentice Hall series in artificial intelligence (2nd ed.). Upper Saddle, New Jersey: Prentice Hall.

S. King, P. Taylor, (2000) "Detection of phonological features in continuous speech using neural networks," *Computer Speech and Language*, *14*, 333-353.

Kirchhoff, K. (1999) *Robust speech recognition using articulatory information*, Ph.D. thesis, University of Bielefield, 1999.

Kirchhoff, K. (1996). Syllable-level desynchronisation of phonetic fea- tures for speech recognition. In *Proceedings of Interspeech* (pp. 2274–2276).

Kirchhoff, K., Fink, G. A. , Sagerer, G. (2002) Combining acoustic and articulatory feature information for robust speech recognition, *Speech Communication*, 37, 303–319.

Kirchhoff, K., (1998) Combining articulatory and acoustic information for speech recognition in noisy and reveberant environments, *Proc. ICSLP*, 1998, pp. 891–894.

Kirchhoff, K., Schimmel, S., (2005). Statistical properties of infant-directed versus adult-directed speech: insights from speech recognition. *Journal of the Acoustical Society of America 117* (4), 2238–2246.

Kleinschmidt, D. F., & Jaeger, T. F. (2014). Robust speech perception : Recognize the familiar , generalize to the similar , and adapt to the novel, *Psychological Review, 122*(2), 148–203. doi:10.1037/a0038695

Kraljic, T., & Samuel, A. G. (2006). Generalization in perceptual learning for speech. *Psychonomic Bulletin & Review*, *13*(2), 262–8.

Lahiri, A. & W.D. Marslen-Wilson (1991) The mental representation of lexical form:a phonological approach to the recognition lexicon. *Cognition 38* , 245-294.

Lahiri, A. & Reetz, H. Distinctive Features: Phonological underspecification in representation and processing. *Journal of Phonetics 38* (2010) 44-59

Lee, L., Rose, R.C., 1998. A frequency warping approach to speaker normalization. IEEE Transactions on Speech and Audio Processing 6(1), 49–60.

Li, J., ; Deng L, ; Gong Y., Haeb-Umbach, R. (2014) An Overview of Noise-Robust Automatic Speech Recognition. *IEEE Transactions on Audio, Speech and Langauge Processing. 22*(4), 745-777.

Lippmann, RP., (1989) Review of neural networks for speech recognition. *Neural computation 1* (1), 1-38

Lippmann, R. P. (1997). Speech recognition by machines and humans. *Speech Communication*, 22(1), 1–15. doi:10.1016/S0167-6393(97)00021-6

Livescu, K., Bezman, A., Borges, N., Yung, L., Çetin, Ö., Frankel, J., King, S., Magimai-Doss, M., Chi, X., Lavoie, L., 2007. Manual transcriptions of conversational speech at the articulatory feature level. In: Proceedings of ICASSP. Vol. 1. pp. 953-956.

Marslen-Wilson, W., & Warren, P. (1994). Levels of perceptual representation and process in lexical access: words, phonemes, and features. *Psychological Review*, *101*(4), 653–75. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/7984710

Mcclelland, J. L., & Elman, J. L. (1986). The TRACE Model of Speech Perception. *Cognitive Psychology*, *18*(1), 1–86.

McClelland, J. L., & Rumelhart, D. E. (1986). *Parallel distributed processing: Explorations in the Microstructure of Cognition*. *(Vol. 2: Psychological and Biological Models)*. Cambridge, MA: MIT Press.

McMillan, R. (2013) How Google Retooled Android with Help from your Brain. Wired Magazine, retrieved from: http://www.wired.com/2013/02/android-neural-network/

McMurray, B., Tanenhaus, M. K., & Aslin, R. N. (2009). Within-category VOT affects recovery from ‘‘lexical” garden-paths: Evidence against phoneme-level inhibition. *Journal of Memory and Language, 60*, 65–91.

Mehler, J., Sebastian-Gallés, N., Altmann, G., Dupoux, E., Christophe, A., & Pallier, C. (1993). Understanding compressed sentences: the role of rhythm and meaning. In A. M. G. Paula Tallal, Rodolfo R. Llinas, Curt von Euler (Ed.), Temporal information processing in the nervous system: Special reference to dyslexia and dysphasia. Annals of the New York Academy of Sciences (Vol. 682, pp. 272-282).

Meyer, B.T., Brand, T., and Kollmeier, B. (2011). "Effect of speech-intrinsic variations on human and automatic recognition of spoken phonemes", J. Acoust. Soc. Am. 129, pp. 388-403. [url | pdf - see copyright notice below (1)]

Meyer, B., Wesker, T., Brand, T., Mertins, A., Kollmeier, B., 2006. A human-machine comparison in speech recognition based on a logatome corpus. Proceedings of the workshop on Speech Recognition and Intrinsic Variation, Toulouse, France.

Miller GA, Heise GA, Lichten W (1951) The intelligibility of speech as a function of the context of the test materials. J Exp Psychol 41:329 –335.

Mirman, D., Estes, K. G., & Magnuson, J. S. (2010). Computational modeling of statistical learning: Effects of Transitional probability versus frequency and links to word learning. *Infancy*, *15*(5), 471–486. doi:10.1111/j.1532-7078.2009.00023.x

Mirman, D., McClelland, J. L., & Holt, L. L. (2006). An interactive Hebbian account of lexically guided tuning of speech perception. *Psychonomic Bulletin & Review*, *13*(6), 958–965. doi:10.3758/BF03213909

Mitterer, H., & McQueen, J. M. (2009). Foreign subtitles help but native-language subtitles harm foreign speech perception. *PloS One*, *4*(11), e7785. doi:10.1371/journal.pone.0007785

Mohamed, A. R., Dahl, G. E. and Hinton, G. E. (2009). Deep belief networks for phone recognition. Proceedings of the Neural Information Processing Systems Workshop on Deep Learning for Speech Recognition

Moore, R.K., 2003. A comparison of the data requirements of automatic speech recognition systems and human listeners. Proceedings of Eurospeech, Geneva, Switzerland, pp. 2581-2584.

Moore, R.K., Cutler, A., 2001. Constraints on theories of human vs. machine recognition of speech. In: Smits, R., Kingston, J., Nearey, T.M., Zondervan, R. (Eds.), Proceedings of the Workshop on Speech Recognition as Pattern Classification. Nijmegen, MPI for Psycholinguistics, pp. 145–150.

Nakatani, L. H., & Dukes, K. D. (1977). Locus of segmental cues for word juncture. *Journal of the Acoustical Society of America*, 62, 715-719.

Norris, D. (1990). A dynamic-net model of human speech recognition. In G. T. M. Altmann (Ed.), *Cognitive models of speech processing*. Cambridge, MA: MIT Press.

Norris, D., McQueen, J. M., & Cutler, A. (2003). Perceptual learning in speech. *Cognitive Psychology*, *47*(2), 204–238. doi:10.1016/S0010-0285(03)00006-9

Norris, D., & McQueen, J. M. (2008). Shortlist B: a Bayesian model of continuous speech recognition. *Psychological Review*, *115*(2), 357–395. doi:10.1037/0033-295X.115.2.357

Ostendorf, M., 1999. Moving beyond the `beads-on-a-string' model of speech. In: Proceedings of IEEE ASRU Workshop. pp. 79-84.

Pearlmutter, B. A. (1995). Gradient calculation for dynamic recurrent neural networks: A survey. IEE Transactions on Neural Networks, 6, 1212–1228.

Puurula, A., & Van Compernolle, D. (2010). Dual stream speech recognition using articulatory syllable models. *International Journal of Speech Technology*, *13*(4), 219–230. doi:10.1007/s10772-010-9080-2

Pinker, S (1994). The Language Instinct: The New Science of Language and Mind. William Morrow.

Remez, R. E., Dubowski, K. R., Broder, R. S., Davids, M. L., Grossman, Y. S., Moskalenko, M., … Hasbun, S. M. (2011). Auditory-phonetic projection and lexical structure in the recognition of sine-wave words. *Journal of Experimental Psychology. Human Perception and Performance*, *37*(3), 968–77. doi:10.1037/a0020734

Robinson, AJ. (1994). An Application of Recurrent Nets to Phone Probability Estimation. IEEE Transactions on Neural Networks, 5(2), 298-305.

Rosenblatt, Frank (1958), The Perceptron: A Probabilistic Model for Information Storage and Organization in the Brain, Psychological Review, v65, No. 6, pp. 386–408

Rumelhart, D.E., Hinton, G, Williams, R. J. (1986) Learning representations by back-propagating errors. Nature, 323, 533-536,

Rumelhart, D. E., & McClelland, J. L. (1986). *Parallel distributed processing: Explorations in the microstructure of cognition*. *(Vol. 1: Foundations)*. Cambridge, Mass: MIT Press.

Salverda, A. P., Dahan, D., & McQueen, J. M. (2003). The role of prosodic boundaries in the resolution of lexical embedding in speech comprehension. *Cognition*, *90*(1), 51–89. doi:10.1016/S0010-0277(03)00139-2

Samuel, A. G., & Kraljic, T. (2009). Perceptual learning in speech perception. *Attention, Perception & Psychophysics*, *71*, 1207–1218.

Scharenborg, O., Norris, D., Bosch, L., & McQueen, J. M. (2005). How should a speech recognizer work? *Cognitive Science*, *29*(6), 867–918. doi:10.1207/s15516709cog0000_37

Scharenborg, O. (2007). Reaching over the gap: A review of efforts to link human and automatic speech recognition research. Speech Communication – Special Issue on Bridging the Gap between Human and Automatic Speech Processing, 49, 336-347.

Scharenborg, O. (2010). Modeling the use of durational information in human spoken-word recognition. Journal of the Acoustical Society of America, 127 (6), 3758-3770.

Scharenborg O., Wan V, Moore R.K. (2007), "Towards capturing fine phonetic variation in speech using articulatory features," *Speech Communication*, 49, 811-826.

Schuppler, B., van Doremalen, J., Scharenborg, O., Cranen, B., Boves, L. (2009). Using temporal information for improving articulatory-acoustic feature classification. Proceedings of the IEEE Automatic Speech Recognition and Understanding Workshop, Merano, Italy, pp. 70-75.

Shatzman, K. B., and McQueen, J. M. (**2006a**). "Segment duration as a cue to word boundaries in spoken-word recognition," Percept. Psychophys. **68**, 1–16.

Shatzman, K. B., and McQueen, J. M. (**2006b**). "The modulation of lexical competition by segment duration," Psychon. Bull. Rev. **13**, 966–971.

Siniscalchi, S. M., Lee, C.-H., 2014. An attribute detection based approach to automatic speech recognition. Loquens 1 (1), http://dx.doi.org/10.3989/loquens.2014.005.

Sohoglu, E., Peelle, J. E., Carlyon, R. P., & Davis, M. H. (2012). Predictive top-down integration of prior knowledge during speech perception. *The Journal of Neuroscience : The Official Journal of the Society for Neuroscience*, *32*(25), 8443–53. doi:10.1523/JNEUROSCI.5069-11.2012

Sroka, J.J., Braida, L.D., 2005. Human and machine consonant recognition. *Speech Communication 45,* 401–423.

Sturm, J., Bakx, I., Cranen, B., Terken, J.M.B. & Wang, F. (2002). Usability evaluation of a Dutch multimodal system for train timetable information. In Manuel Gonzáles Rodríguez & Carmen Suárez Araujo (Eds.), Proceedings LREC 2002. Third International Conference on Language Resources and Evaluation (pp. 255-261). Paris: ELRA, European Language Resources Association.

Vincent, E., Barker, J., Watanabe, S., Le Roux, J., Nesta, F., & Matassoni, M. (2013). The second "CHiME" speech separation and recognition challenge: An overview of challenge systems and outcomes. *2013 IEEE Workshop on Automatic Speech Recognition and Understanding, ASRU 2013 - Proceedings*, 162–167. doi:10.1109/ASRU.2013.6707723

A. Waibel, T. Hanazawa, G. Hinton, K. Shikano, and K. Lang, "Phoneme recognition using time-delay neural networks," IEEE , Trans. Acoust., Speech, Signal Process., vol. 37, no. 3, pp. 328–339, .

Wester, M., 2003. Syllable classification using articulatory-acoustic features. In: Proceedings of Eurospeech, Geneva, Switzerland, pp. 233–236.

Woodland, P. C. (2001) Speaker Adaptation for Continuous Density HMMs: A Review. In: *Proceedings ISCA Workshop on Adaptation Methods for Speech Recognition, 2001,* 11–19

Young, S.J. (1996). A review of large-vocabulary continuous-speech recognition. *IEEE Signal Processing Magazine, 13* (5), pp. 45-57

Yu, K., & Gales, M. J. F. (2007). Bayesian adaptive inference and adaptive training. *IEEE Transactions on Audio, Speech and Language Processing*, *15*(6), 1932–1943. doi:10.1109/TASL.2007.901300

Zhang, X., Samuel, A.G. (2014) Perceptual learning of speech under optimal and adverse conditions. *Journal of Experimental Psychology: Human Perception and Performance, 40*(1), 200-217.