ScienceDirect

Automatic Speech Recognition

Automatic speech recognition (ASR) is the process and the related technology for converting the speech signal into its corresponding sequence of words or other linguistic entities by means of algorithms implemented in a device, a computer, or computer clusters (Deng and O'Shaughnessy, 2003; Huang et al., 2001b).

From: Robust Automatic Speech Recognition, 2016

Related terms:

Speech Recognition, Deep Neural Network, Application Domain, Microphone, Audio Feature, Audio Segmentation, Audio Signal, Phase Space, Speech Signal

Introduction

Jinyu Li, ... Yifan Gong, in Robust Automatic Speech Recognition, 2016

Abstract

Automatic speech recognition (ASR) by machine has been a field of research for more than 60 years. The industry has developed a broad range of commercial products where ASR as user interface has become ever more useful and pervasive. Consumer-centric applications increasingly require ASR to be robust to the full range of real-world noise and other acoustic distorting conditions. However, reliably recognizing spoken words in realistic acoustic environments is still a challenge.

We introduce distortion factors that operate in various stages of speech production, from thought to speech signals, leading to the issues of ASR robustness as the focus of this book. We provide an introductory summary of this book in this chapter, covering the ASR robustness problem for acoustic models based on both Gaussian mixture models and deep neural networks. The book goes significantly beyond much of the existing survey literature, and illustrates the research and product development on ASR robustness to noisy acoustic environments that has been progressing for over 30 years.

Finally, we define the mission, goal, and structure of the book in this chapter. We aim to establish a solid, consistent, and common mathematical foundation for robust ASR, emphasizing the methods proven to be successful and expected to sustain or expand their future applicability.

The State of the Art in Human–Robot Interaction for Household Services

Dan Xu, ... Yangsheng Xu, in Household Service Robotics, 2015

6.1.1.2 Speech-Based HRI Systems

With the rapid progress of automatic speech-recognition techniques [31–34], speech-based human–robot interaction (sHRI) has attracted increasing attention from the robotics research community. The researchers have developed many speech-based HRI systems that cover a wide range of application scenarios, and we briefly introduce several of these in this subsection.

To provide more robust models of language understanding for natural HRI, Cantrell et al. described an integral natural language understanding architecture for HRI [35], and the capabilities of the system were demonstrated through two experiments on the spoken instruction understanding of robots, including semantic ambiguities and incremental understanding with back-channel feedback.

Breazeal and Aryananda proposed an approach to recognize four distinct prosodic patterns to represent communication intent, including praise, prohibition, attention, and comfort, to preverbal infants, and this approach was integrated into the “emotion” system of a robot to enable humans to directly manipulate the robot's affective states [36].

For affective robot–child interaction, expressive speech synthesis and recognition are considered enabling techniques. A new speech synthesizer was developed by Yimazyildiz et al. [37] to allow the robot to synthesize expressive nonsense speech and then work toward effective affective interactions with children.

Kriz et al. developed speech-based HRI systems based on robot-directed speech to study the conceptualizations of robots [38,39]. Shown in Figure 3, is an experimental scenario in which the human asks the robot to fetch certain objects through robot-directed speech.

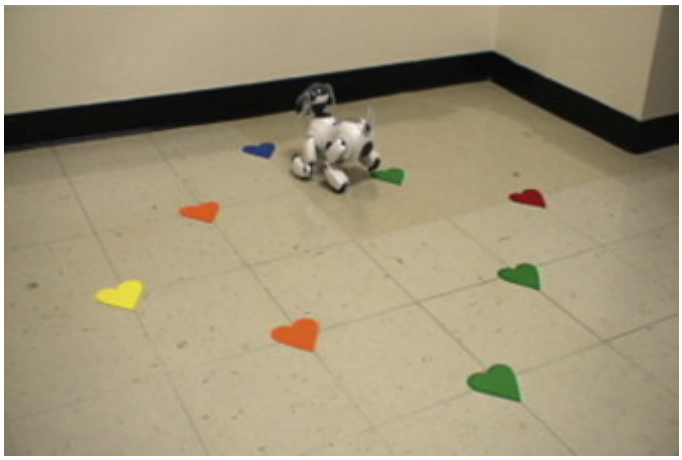

Figure 3. Speech-based interactions with AIBO.

In RoboCup 2008, Doostdar et al. proposed a speaker-independent speech-recognition system [40], using off-the-shelf technology and simple additional approaches, which can obtain high recognition accuracy under experimental conditions of loud noise and meets the needs of the mobile-service robot working

in human environments.

Speech Communication: How to Use It

John Waterworth, in Fundamentals of Human–Computer Interaction, 1985

13.3 VOICE RECOGNITION

The use of speech recognition equipment (Automatic Speech Recognition or ASR) has several advantages over keyed input, principally because command words can be selected to match the functions they initiate, so that there is no need for the user to map functions onto arbitrary symbols. But the available technology does not, at present, lend itself to particularly comfortable interactions between man and machine, despite the claimed naturalness and convenience of the medium. Dialogue design must take account of the fact that the word 'recognised' will often be the wrong one, so that the user will frequently need to confirm or correct the input the machine has identified.

Measuring Recognition Accuracy

Choosing the most suitable recogniser for a particular task can be fraught with difficulties. Manufacturers typically quote figures of between 95 and 99% accuracy for their products, of whatever type. But without considering other information this is largely meaningless. Such figures are presumably derived by very carefully selecting a small, easily distinguishable vocabulary, and testing with experienced users in ideal conditions. Accuracy of recognition (in terms of the number of hits out of the number of attempts) is important, but the circumstances under which scores are obtained is crucial, and accuracy is not the whole story; confusions, too, need to be taken into account. We have found that confusion matrices and trees are useful in selecting task vocabularies for particular applications. A confusion tree of the digits 1 to 5 and five command words, obtained from an inexpensive single-word recogniser, is illustrated in Figure 13.1. The results shown are from 60 subjects, each repeating all ten words 20 times (i.e., 1200 utterances of each word). Certain words are very frequently confused (e.g., '1' and '5', and '3' and 'REPEAT'), whereas for others this is hardly ever the case (e.g., '1' and 'REPEAT').

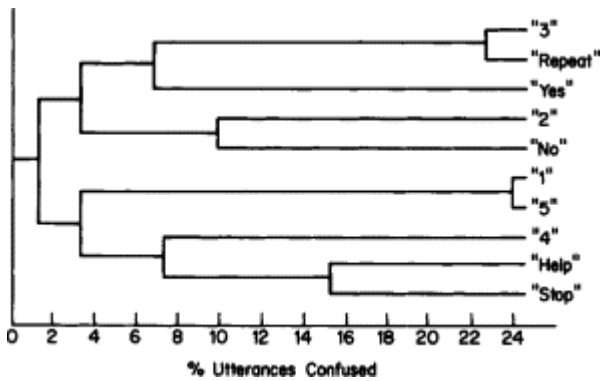

FIG. 13.1. Confusion tree for the digits 1 to 5 and five command words. Items linked at the right hand side of the tree are easily confused (e.g. "1" and "5"); items linked at the left hand side are not easily confused (e.g. "No" and "1").

What the confusion tree does not show, although a confusion matrix would, is that

the pattern is often not symmetrical. In the same trials, 'REPEAT' was misrecognised as '3' over three times more frequently than '3' was misrecognised as 'REPEAT', for example. Figure 13.2 presents a confusion tree for the digits 1 to 10 from 24 subjects, each repeating all ten words 32 times (i.e., 768 utterances of each word). This clearly indicates that certain digits, notably '5' and '9', should not be used together with this device, if at all possible. If it proved necessary to recognise all the digits, then users would have to be encouraged to mispronounce the words to make them more distinguishable (as with the "fife" and "niner" of Civil Aviation English). Consideration of the probable pattern of vocabulary confusions is essential to successful command word selection for a given application.

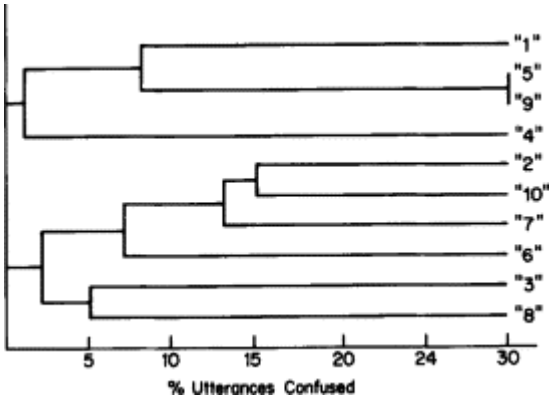

FIG. 13.2. Confusion tree for the digits 1 to 10.

Recognisers really need to be evaluated on the intended vocabulary, perhaps using a procedure analogous to that described for assessing speech synthesisers (see Chapter 7). Even better is testing the machine in the target application, but this cannot be done in advance of acquiring the device. An approach that is becoming popular is to provide standard vocabulary tapes for a range of different types of machine. Manufacturers will inevitably tailor their devices to perform well on the test tapes, however, so that the claimed performance may not be achieved in general use unless the material on the tapes is very carefully chosen.

The Limitations Imposed by Existing Systems

Currently available recognition systems impose a large number of constraints on the forms the man-machine interaction can be allowed to take. For example, they are usually only capable of recognising words spoken in isolation and often only by a speaker who has previously 'trained' the system on a particular and very limited vocabulary. With isolated word recognition, the system must interrogate the user in a way that encourages a 'legal' responding utterance. This means that careful consideration must be given to the design of the man-machine speech dialogue, to elicit appropriate responses from the user.

Many devices require the user to repeat the vocabulary of words to be recognised several times in order to 'train' the recogniser. The time and effort required to train a recogniser to a particular voice are likely greatly to inhibit the use of a system. It has been suggested that disguising the training session elicits utterances more similar to those produced during the actual task than when a formal training session is used, although the success of this approach was not tested. The method depends on the user's utterances being predictable, while the user gets the

impression that he is actually deciding what to say. For the user, the distinction between training and use is supposed to disappear. While this may avoid some of the irritation generated by the training session, the extent to which the hidden training truly resembles the task situation, from the user's point of view, is more doubtful. It is very difficult to think of training tasks that realistically mimic the actual application, give the impression of subjective control, while being, in fact, totally determined. Such training tasks tend to be trivially easy for users, unlike actually using the system. So the claimed advantage of this technque, that similar utterances are produced in training and actual use, may be more apparent than real. Voice prompting also has implications in this area. We have observed that users who are training a recogniser in response to synthetically produced vocal prompts have a tendency to imitate the way in which the prompts are pronounced. If this is also true during actual use of the system, recognition might be enhanced. But if the user does not also imitate the prompting voice during use of the service recognition performance will be degraded. The effects of visual versus auditory prompting need to be tested. The use of a very brief training session, to fine-tune preprogrammed templates to a particular user, is likely to increase the acceptability of 'user-dependent' devices in the future.

For a limited set of regular, motivated users, the need to train the system, and occasionally update templates, is not a major problem. Of course, the ultimate goal must be to develop techniques whereby no training of the system by the user is required, and this is particularly important for telephone-accessible services used by the general public. The user cannot be expected to train the system on his voice every time he wants to know the time of a bus or a weather report. Recognisers that are available now, and claim to be speaker independent, have actually been trained by a large set of users to produce several versions of each word template, or merge many utterances into composite templates. Unfortunately, neither approach is very satisfactory, and the best way of producing general-user templates is not known. British Telecom are currently investigating one approach to this problem, that of using different voice 'type' exemplars as templates, from which a particular set could be selected for each user on the basis of his first utterances. However, it is unlikely that any one system would be able to cope with the full range of British accents and dialects, at least for the foreseeable future. And there will always be a subset of the population for whom automatic speech recognition is not possible, hence the need for operator intervention in cases of serious difficulty.

Connected word recognition is now available, although this usually still involves training on discrete single words. The problems associated with using a recogniser can actually increase with connected-word recognition, because natural connected speech does not consist of a sequence of complete, nonoverlapping whole words. Of course, the user will only be successful if he uses words on which the system has been trained, and he needs to speak in a very artificial way. It is difficult to find a simple way of encouraging the user to make acceptable responses. It is also much harder to give the user a helpful model of how the system operates than is the case with single-word machines. The concept of a speech recogniser that recognises all of some phrases, parts of others, and some not at all, is far from obvious. Another major disadvantage is that this approach is not practically extensible to continuous speech recognition from unconstrained discourse, because of lexical demands.

Systems Design for Automated Speech Recognition

Martin Helander, ... Michael G. Joost, in Handbook of Human-Computer Interaction, 1988

Publisher Summary

This chapter discusses a systems design for automated speech recognition. Automatic speech recognition (ASR) is a method for communicating with or inputting data into a computer. The user talks into a microphone and the characteristic features of individual words or utterances are extended using an algorithm that is resident in the computer. The features are then compared to a library of characteristics that are stored in computer memory, and the best match is selected as the preferred word. The intuitive appeal of speech recognition is that speech is the most natural way of communicating and is, for this reason, preferred over other types of input. In fact, it has the potential of being an ideal input device, especially for novice users and when combined with other input modalities. However, using ASR is not as simple or effortless as talking to a person. The act of choosing words from a restricted vocabulary and speaking in a precise and consistent fashion requires much conscious effort and attention.

Fundamentals of speech recognition

Jinyu Li, ... Yifan Gong, in Robust Automatic Speech Recognition, 2016

Abstract

In this chapter, we introduced the fundamental concepts and component technologies for automatic speech recognition. The topics reviewed in this chapter include several important types of acoustic models—Gaussian mixture models (GMM), hidden Markov models (HMM), and deep neural networks (DNN), plus several of their major variants. The role of language modeling is also briefly discussed in the context of the fundamental formulation of the speech recognition problem.

The HMM with GMMs as its statistical distributions given a state is a shallow generative model for speech feature sequences. Hidden dynamic models generalize the HMM by incorporating some deep structure of speech generation as the internal representations of speech features. Much of the robust speech recognition studies in the past were carried out based on generative models of speech, since the noisy version of speech as the observation signal can be easily "generated" from clean speech using straightforward distortion models. Recently, the discriminative DNN, as well as its convolutional and recurrent variants, have been shown to significantly outperform all previous versions of generative models of speech in speech recognition. The main classes of these deep discriminative models are reviewed in some detail in this chapter. How to handle noise robustness within the framework of discriminative deep learning models of speech, which is less straightforward than the generative models of speech, will be covered in the later chapters of this book.

Summary and future directions

Jinyu Li, ... Yifan Gong, in Robust Automatic Speech Recognition, 2016

In this book, we first presented fundamental concepts, components, models, and

methods of automatic speech recognition (ASR). This introductory material included Gaussian mixture models (GMM), hidden Markov models (HMM) and variants, and the more recent deep neural networks (DNN) and related deep learning methods. We then provided background information about the robust ASR problem based on the GMM-HMM and DNN speech models. One most important concept in the robust ASR problem is acoustic distortion of speech signals and features. The effects of such acoustic distortion on both GMM-HMM and DNN models of speech are elaborated in a mathematically rigorous manner. To offer insight into the distinct capabilities of a wide range of noise-robust ASR techniques combating the acoustic distortion, we use a taxonomy-oriented approach for single-microphone nonreverberant speech recognition. The taxonomy adopted is based on five key attributes—feature vs. model domain processing, explicit vs. implicit distortion modeling, use of prior knowledge about distortion or otherwise, deterministic vs. uncertain processing, and joint vs. disjoint training. These attributes are used to organize the literature on the topic spanning over more than three decades and to demonstrate the commonalities and differences among the plethora of robust ASR methods described in this book. Each of the five attributes constitutes a separate chapter in the book. Then after the noise-robust techniques for single-microphone nonreverberant ASR are comprehensively discussed, we include two chapters, covering reverberant ASR and multi-channel processing for noise-robust ASR, respectively.

Advances in Computers: Improving the Web

Dalibor Mitrović, ... Christian Breiteneder, in Advances in Computers, 2010

2.1 A Brief Overview on Content-Based Audio Retrieval

There are different fields of research in content-based audio retrieval, such as segmentation, automatic speech recognition, music information retrieval, and environmental sound retrieval which we list in the following. *Segmentation* covers the distinction of different types of sound such as speech, music, silence, and environmental sounds. Segmentation is an important preprocessing step used to identify homogeneous parts in an audio stream. Based on segmentation, the different audio types are further analyzed by appropriate techniques.

Traditionally, *automatic speech recognition* focuses on the recognition of the spoken word on the syntactical level [1]. Additionally, research addresses the recognition of the spoken language, the speaker, and the extraction of emotions.

In the last decade *music information retrieval* became a popular domain [2]. It deals with retrieval of similar pieces of music, instruments, artists, musical genres, and the analysis of musical structures. Another focus is music transcription which aims at extracting pitch, attack, duration, and signal source of each sound in a piece of music [3].

Environmental sound retrieval comprises all types of sound that are neither speech nor music. Since this domain is arbitrary in size, most investigations are restricted to a limited domain of sounds. A survey of techniques for feature extraction and classification in the context of environmental sounds is given in Ref. [4].

One major goal of content-based audio retrieval is the identification of perceptually similar audio content. This task is often trivial for humans due to powerful mechanisms in our brain. The human brain has the ability to distinguish between a

wide range of sounds and to correctly assign them to semantic categories and previously heard sounds. This is much more difficult for computer systems, where an audio signal is simply represented by a numeric series of samples without any semantic meaning.

Content-based audio retrieval is an ill-posed problem (also known as inverse problem). In general, an ill-posed problem is concerned with the estimation of model parameters by the manipulation of observed data. In case of a retrieval task, model parameters are terms, properties, and concepts that may represent class labels (e.g., terms like “car” and “cat,” properties like “male” and “female,” and concepts like “outdoor” and “indoor”).

The ill-posed nature of content-based retrieval introduces a *semantic gap*. The semantic gap refers to the mismatch between high-level concepts and low-level descriptions. In content-based retrieval, the semantic gap is positioned between the audio signals and the semantics of their contents. It refers to the fact that the same media object may represent several concepts. For example, a recording of Beethoven's Symphony No. 9 is a series of numeric values (samples) for a computer system. On a higher semantic level the symphony is a sequence of notes with specific durations. A human may perceive high-level semantic concepts like musical entities (motifs, themes, movements) and emotions (excitement, euphoria).

Humans bridge the semantic gap based on prior knowledge and (cultural) context. Machines are usually not able to complete this task. Today, the goal of the research community is to narrow the semantic gap as far as possible.

Predicting Short-Term Congested Traffic Flow on Urban Motorway Networks

Taiwo Adetiloye, Anjali Awasthi, in Handbook of Neural Computation, 2017

8.3.3 Deep Learning

Since 2006, deep learning has evolved as a class of machine learning methods with successful applications in various fields like automatic speech recognition, classification tasks, natural language processing, dimensionality reduction, object detection, motion modeling, etc. [6,11,22,25,28]. Its algorithms are based on the architecture of hierarchical explanatory factors and distribution representations where a cascade of many layers of nonlinear processing units is used for the supervised or unsupervised learning of feature representations per layer, with the layers forming a hierarchy from low-level to high-level features, in the sense of feature extraction and transformation [14]. It is inspired by some loosely established interpretation of information systems and communication patterns formulated on the human nervous system, which attempts to model high-level abstractions in data using a deep graph with multiple processing layers. Some notable architectures of deep learning include the deep belief networks, convolutional neural networks, and recurrent neural networks [5,6,23,54]. A detailed discussion of deep learning in neural networks can be found in Schmidhuber [54].

In traffic flow prediction, the use of deep learning approaches for features extraction and selection without any prior knowledge has been investigated by

Huang et al. [27]. Its stack architecture used for traffic flow prediction of a single road is comprised of Restricted Boltzmann Machines (RBM) stacks constituting the DBN for the unsupervised feature learning having sigmoid regression layer on top. This came out of concern about the shallow architectures of neural network attributed to the single hidden layer, hence introducing the key idea of using greedy layer-wise training with stacked RBM and subsequent fine tuning according to the architecture of the DBN guaranteed near 3% improvements over state-of-the-art. A deep learning approach for traffic flow prediction with big data was introduced by Lv et al. [43] as a means for accurate and timely traffic flow information while considering the spatiotemporal correlation present in an intelligent transportation system. The following subsection covers the method using DBN.

Deep Belief Networks In machine learning, DBN is a multilayered probability generative model composed of simple learning modules, so-called RBMs [23], also known as autoencoders [5], where each subnetwork's hidden layer serves as the visible layer for the next [5,24]. An RBM implies the absence of the lateral connections in the visible and hidden layers such that the random variables encoded by the hidden units are conditionally independent given the states of the visible units, and vice versa [46]. In Teh and Hinton [60], RBM is defined as "an undirected graphical model in which visible variables (*v*) are connected to stochastic hidden units (*h*) using undirected weighted connections." The architecture of DBN can be much more efficient than shallow architectures as contained in the single latent layer of feedforward and BP-NN with many levels of non-linearity and highly-varying functions. Training the deep layer networks one layer at a time using greedy algorithm, rather than the gradient-based optimization often used with BP-NN, can guarantee a good local optimum [5], though with the BP-NN, feasible solution can be reached if starting in the neighborhood of a good local optimum. Also, BP-NN is liable to poor performance and prone to overfitting [26]. Moreover, despite the number of NNs that exist, finding one that precisely model a given training set can be an N-P complete problem [7,31,54]. It is noteworthy that with the labels provided, the DBM could provide a supervised feature learning model suitable for classification. While, considering the DBN in the classification tasks of traffic congestion with the binary RBM, we assigned the visible input unit, v_i, as elements of the encoded joint distribution function. Based on Hinton and Sejnowski [22] and as further illustrated by O'Connor et al. [46], the encoded joint distribution function can be defined as

$$p(v,h\,|\,\theta)=\frac{\exp(-E(v,h;\theta))}{\sum_{v'}\sum_{h'}\exp(-E(v',h';\theta))} \tag{8.11}$$

where the energy function is given by:

$$E(v,h;\theta)=-\sum_i\sum_j w_{ij}v_ih_j-\sum_i b_i^{(v)}v_i-\sum_j b_j^{(h)}v_j \tag{8.12}$$

With the model parameters, $\theta=(w,b^{(v)})$, h_j is the states of the hidden units, w_{ij} represents the states connecting these units: namely, the visible input and hidden units; and $b_i^{(v)}$ and $b_j^{(h)}$ are the biases in the visible and hidden units respectively. Fig. 8.4 illustrates the DBN architecture.

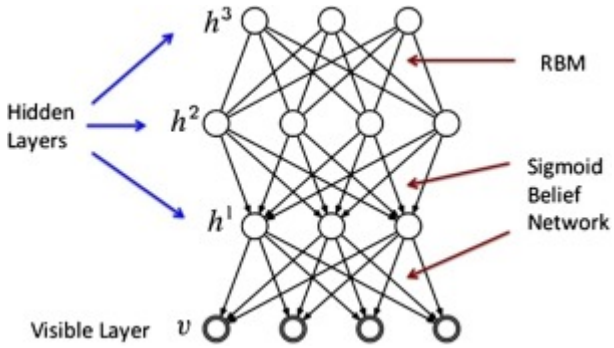

Figure 8.4. Deep belief network.

Source: Source: Hinton et al. [24]

An effective way to learn θ using contrastive divergence has been proposed by Hinton et al. [24]. Hence, we adapt the method that employs a greedy layer-wise algorithm associated with the RBM to the traffic data with the important factors: average speed, traffic flow, and the link-length assigned to the input units. The model establish the following stochastic update rules for the state for which lower energy state eventually attains an equilibrium given by:

$$p(v_i = 1, h \mid \theta) = \sigma\left(\sum_j w_{ij} h_j + b_i^{(v)}\right) \tag{8.13}$$

$$p(h_j = 1, h \mid \theta) = \sigma\left(\sum_j w_{ij} v_i + b_j^{(h)}\right) \tag{8.14}$$

where $\sigma(t) = \frac{1}{1+e^t}$ denotes the sigmoid function with the capability of having its unit state change to 0 and the flexibility of the network to generate samples over all possible states (v, h) based on the joint probability distribution, $p(v, h \mid \theta)$.

Multi-channel processing

Jinyu Li, ... Yifan Gong, in Robust Automatic Speech Recognition, 2016

10.6 Summary

Table 10.1 summarizes key publications in the field of multi-channel processing. We mention only very few from the field of beamforming, being aware of the fact that there are many more influential publications in this field. Acoustic beamforming for automatic speech recognition was pioneered in the 1990s by Compernolle and others. Multi-Stream ASR became popular in the late 1990s with works by Bourlard and Morgan, among others. Seltzer et al. investigated a closer coupling between acoustic beamforming and speech recognition. They optimized the beamformer coefficients using criteria relevant to ASR. Recently, various approaches to combine multi-channel data and deep neural networks have been investigated.

Table 10.1. Approaches to Speech Recognition in the Presence of Multi-Channel Recordings

Method	Proposed Around	Characteristics
MVDR beamformer (Frost, 1972)	1972	Minimizes output power of the beamformer under a constraint in the look direction
Generalized sidelobe cancellor (Griffiths and Jim, 1982)	1982	Reformulates MVDR criterion as an unconstrained optimization
Speech recognition with microphone arrays (Compernolle et al., 1990; Omologo et al., 1997)	1990-1997	First publications demonstrating the benefits of beamforming for ASR
Multi-stream ASR (Bourlard et al., 1996; Janin et al., 1999)	1996	Multiple input streams to ASR with fusion at feature, score, or decoding stage
Concept of relative transfer functions (Cohen, 2004; Gannot et al., 2001; Warsitz and Haeb-Umbach, 2007)	2001	Different methods to identify relative transfer functions
Multi-channel Wiener filter (Doclo and Moonen, 2002)	2002	Time domain realization of (parametric) multi-channel Wiener filter
Likelihood maximizing beamforming (Seltzer et al., 2004)	2004	Determines beamformer coefficients as to maximize the likelihood of the correct word hypothesis in the recognizer
Beamforming criterion based on higher-order statistics (Kumatani et al., 2009)	2009	Maximum negentropy criterion for beamformer design
Convolutional neural networks for multi-channel input (Swietojanski et al., 2014)	2014	Convolutional neural networks for distant speech recognition

Example of an Experiment: Evaluating Some Speech Synthesisers for Public Announcements

John Waterworth, Antony Lo, in Fundamentals of Human–Computer Interaction, 1985

Background

British Telecom are involved in a broad range of human factors research concerned with making the process of man-machine interaction as successful and effortless for users as possible. Access to information-providing systems by means of speech-based interaction is of particular interest to us, because this permits users to

communicate directly with information bases by means of an ordinary telephone. Keyed input or automatic speech recognition, combined with synthetic speech output from the computer, provide the means of communication. These services are principally aimed at the general public, so that no expertise in computer interactions can be assumed to be held by users. This means that very close attention must be directed towards 'the user interface'. We have conducted several experimental studies to maximise the effectiveness of voice-based interactive services. These have included the design and evaluation of complete systems (actual existing systems and simulated services), as well as more specific studies on particular aspects of speech synthesis and recognition. Chapter 13 provides a review of some of the work in these areas.

This chapter describes two experiments that examined the use of synthetic speech in the context of an interactive train timetable service. The first compared the performance of several different synthesisers, while the second focussed on the effects of synthesiser speaking rate.