Towards Cognizant Hearing Aids: Modeling of Content, Affect and Attention

Karadogan, Seliz Gülzen; Larsen, Jan

Publication date:
2012

Document Version
Publisher's PDF, also known as Version of record

Towards Cognizant Hearing Aids: Modeling of Content, Affect and Attention

Seliz Gülsen Karadoğan

Kongens Lyngby 2012
IMM-PhD-2012-275

Technical University of Denmark
Informatics and Mathematical Modelling
Building 321, DK-2800 Kongens Lyngby, Denmark
Phone +45 45253351, Fax +45 45882673
reception@imm.dtu.dk
www.imm.dtu.dk IMM-PhD-2012-275

Summary (English)

Hearing aids improved significantly after the integration of advanced digital signal processing applications. This improvement will continue and evolve through obtaining intelligent, individualized hearing aids integrating top-down (knowledge-based) and bottom-up (signal-based) approaches by making use of research done within cognitive science that is the interdisciplinary study of mind and intelligence bringing together various disciplines including Artificial Intelligence, Cognitive Psychology, and Neuroscience.

This thesis focuses on three subjects within cognitive science related to hearing. Initially, a novel method for automatic speech recognition using binary features from binary masks, is discussed. The performance of binary features in terms of robustness to noise is compared with the ASR state of the art features, mel frequency cepstral coefficients. Secondly, human top-down auditory attention is studied. A computational top-down attention model is presented and behavioral experiments are carried out to investigate the role of top-down task driven attention in the cocktail party problem. Finally, automatic emotion recognition from speech is studied using a dimensional approach and with a focus of integrating semantic and acoustic features. An emotional speech corpus that consists of short movie clips with audio and text parts, rated by human subjects in two affective dimensions (arousal and valence), is prepared to evaluate the method proposed.

Summary (Danish)

Integrationen af avanceret digital signalbehandling har været en betydelig innovation indenfor høreapparatsindustrien. Denne udvikling vil fortsætte gennem intelligente og individualiserede høreapparater der integrerer top-down (videnbaseret) og bottom-up (signalbaseret) tilgange ved brug af forskning inden for kognitive systemer, som er det tværfaglige studiet af sind og intelligens og samler forskellige discipliner, herunder kunstig intelligens, machine learning, kognitiv psykologi og neurovidenskab.

Denne afhandling fokuserer påtre emner inden for kognitive systemer relateret til hørelse. Indledningsvis foreslås en ny fremgangsmåde til automatisk talegenkendelse ved brug af binære maske features. De binære features robusthed over for støjer sammenlignet med state-of-the-art standard features, nemlig mel- frekvens-kepstrale-koefficienter. Det andet emne som behandles er top-down auditiv opmærksomhed. En top-down opmærksomhed model præsenteres og verificeres gennem adfærdseksperimenter. Bl.a. undersøges betydningen af top-down opgave-drevet opmærksomhed i forbindelse med det velkendte cocktailparty problem. Endelig er automatisk følelsesgenkendelse i tale studeret ved med fokus påat integrere semantisk og akustisk information. Den foreslåede metode er evalueret ved brug af et nyt datasæt, der består af korte filmklip med lyd og tekst dele, som er bedømt af forsøgspersoner i to affektive dimensioner (arousal og valens).

Preface

This thesis was prepared at the department of Informatics and Mathematical Modeling at the Technical University of Denmark (DTU) in fulfillment of the requirements for acquiring a PhD degree in Informatics. The project was supervised by Assoc. Professor Jan Larsen (DTU).

The thesis deals with future hearing aids that are intelligent and personalized taking into account cognitive processing in the human auditory system that we call 'cognizant hearing aids'. The thesis focuses on three subjects within cognitive science that are automatic speech recognition, top-down auditory attention and automatic emotion recognition from speech. This thesis represents a summary of my research with detailed descriptions provided in the associated appendix.

The thesis consists of a summary report and a collection of six research papers written during the period 2009-2012.

Lyngby, 31-May-2012

Seliz Gülsen Karadoğan

Publications included

A S.G. Karadoğan, J. Larsen, M.S. Pedersen, and J.B. Boldt. Robust isolated speech recognition using binary masks. *Proceedings of European Signal Processing Conference (EUSIPCO)*, pp. 1988-1992, 2010.

B S.G. Karadoğan , L. Marchegiani, L.K. Hansen and J. Larsen. How efficient is estimation with missing data? *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2260-2263, 2011.

C L.K. Hansen, S.G. Karadoğan and L. Marchegiani. What to measure next to improve decision making? On top-down task driven feature saliency. *IEEE Symposium Series on Computational Intelligence (SSCI)*, pp. 1-7, 2011.

D S.G. Karadoğan , L. Marchegiani, J. Larsen and L.K. Hansen. Top-down attention with features missing at random. *IEEE International Workshop on Machine Learning For Signal Processing*, pp. 1-6, 2011.

E L.Marchegiani , S.G. Karadoğan , T. Andersen, J.Larsen and L. K. Hansen. The Role of Top-Down Attention in the Cocktail Party: Revisiting Cherry's Experiment after Sixty Years. *International Conference on Machine Learning and Applications (ICMLA)*, pp. 183-188, 2011.

F S.G. Karadoğan, J. Larsen. Dimensional Emotion Recognition from

Speech Combining Semantic and Acoustic Features. *IEEE Transaction on Affective Computing*, submitted, 2012.

Acknowledgments

I would like to thank my supervisor Assoc. Prof. Jan Larsen, first of all, for his help in preparing the PhD project proposal to get the scholarship, for having very critical eyes even for little details and for always encouraging me for managing my own project. I would also like to thank to Prof. Lars Kai Hansen for his inspirational ideas and for his encouraging comments and feedbacks. I would like to thank to Tobias Andersen as well for his insights to some parts of my research.

A special thanks to the review committee, Prof. Søren Holdt Jensen, Assoc. Prof. Ole Winther and DI Dr. Björn W. Schuller, for their time, perspectives and comments towards improving my thesis.

I would like to thank to my colleagues at DTU Informatics for an inspiring environment and for being courageous to volunteer our experiments :), in particular to Michael Kai Petersen also for helping me with some details in my research. I also would like to thank to Arek and Kristian for a fun and 'hyggeligt' office.

I would like to thank to my office, project, conference, summer school, Friday bar, movie nights, hotel room, shopping partner, my neighbor and my confidant Letizia for being such a nice friend, for always having time to listening to my problems during my PhD (not necessarily technical :D) and for always smiling and making me smile. When Lars and Jan

first introduced me to her as the 'new girl' in the group, I was so happy finally to have a girlfriend at work (one of the challenges to be a woman and an engineer :)). Although I thought she did not speak English for the first few days, after being have to be her 'Seliz Translate' for some time and having to use some kind of body language, we have been a 'great' couple since then. I would also like to thank to the DTU (mostly IMM) crew (Letizia, Massimo, Valentina, Fontas, Laura, Dimitri, Fabrice and many more) for the funny Fridays, Friday bars, barbecues, beach days and many more fun events.

I would like to thank to Enis, for his critics on my research, for always having an answer for me or leading me to the answer with his questions when I am confused during my research for which I and Letizia call him 'Enispedia' :), for being my 'bodyguard' :D at MLSP conference in China and for many others that helped me through my PhD.

I would also like to thank to my family, my parents Arife and Nazım for having faith in my decisions and for supporting me, my sweet sister Şeniz Nurten for always reminding me how much she cares about me and to the coolest brother Deniz Ertan for being there for me even from far at my hardest periods during my education.

Abbreviations

A	Arousal
AI	Artificial intelligence
ANEW	The affective norms for English words
ASA	Auditory scene analysis
ASR	Automatic speech recognition
BTA	Both text and audio
CEM	Complete expectation maximization
CFS	Correlation feature selection
CIBM	compressed ideal binary mask
DIR	directed attention experiments
DSP	Digital signal processing
GDMM	Gaussian discrete mixture model
GMM	Gaussian mixture model
HH	High arousal high valence
HI	Hearing impaired

HL	High arousal low valence
HMM	Hidden Markov models
IBM	Ideal binary masks
JA	Just audio
JT	Just text
LC	Local criteria
LH	Low arousal high valence
LL	Low arousal low valence
MAR	Missing at random
MCAR	Missing completely at random
MDT	Missing data techniques
MFCC	Mel frequency cepstral coefficients
MNAR	Missing not at random
MSE	Mean Square Error
NH	Normal hearing
NNSC	Non-negative sparse coding
RC	Relative criteria
RMSE	Root mean squared error
SAM	The self assessment manikin
SNR	Signal to noise ratio
SSN	Speech shaped noise
T-F	Time-frequency
TBM	Target binary masks
UNDIR	undirected attention experiments

V	Valence
WDRC	Wide dynamic range compression

Contents

Chapter 1

Introduction

Hearing aid technology has progressed drastically over the last fifteen years primarily due to the introduction of advanced digital signal processing (DSP) into hearing aids in the late 1990s as a result of progress in the semiconductor technology [Sch08, HCE+05]. The use of DSP has led to many improvements on minimizing the unwanted effect of background noise and acoustic feedback that is caused by the undesired acoustic coupling between the loudspeaker and the microphone [NvWC+10, FWG02]. Directional hearing aids [Ric05], with directional microphones designed to attenuate the sounds coming from an angle other than in front of the listener, were first used in hearing aids in late 1970s, yet gained more importance with the help of DSP that is used to calibrate microphones, control the shape of the directional pattern, automatically switch between directional and omni-directional modes, and through expansion, reduce additional circuit noise generated by directional microphones. Modern hearing aids also offer multi-channel wide dynamic range compression (WDRC) [KRD+06], to increase audibility of weak sounds, and noise reduction schemes [BM09], to increase comfort.

In hearing aid consumer satisfaction studies carried out between 1991

and 2008 [Koc10], it is reported that the satisfaction rate increased from 66.4% in 1991 to 80.6% in 2008 as can be seen in Table 1.1. The increase in satisfaction rate in 2000s (14.4% increase just from 2000 to 2004) is impressive which can also be interpreted as the success of introducing advanced DSP. As DSP was a revolutinary breakthrough in hearing aid industry, there is still much room to advance hearing aids that would result in 100% user satisfaction, enabling hearing impaired (HI) people hear as good as or even better than normal hearing (NH) people, and possibly resulting in NH people to wear them as well. Those incremental innovations are thought to be driven by advances in wireless technology, digital chip technology, hearing science and cognitive science [Edw07].

Satisfaction with hearing aids (</= 1 year)	1991	1994	1997	2000	2004	2008
%Satisfied	66.4%	71.8%	62.9%	63.1%	77.5%	80.6%

Table 1.1: Satisfaction with hearing aids (not older than 1 year) survey results 1991-2008 (taken from [Koc10])

The advances in wireless technology leads to the ability for two hearing aids to communicate directly and continuously with each other and to external audio and communication devices [KCK+11]. Digital wireless transmission makes it possible to exchange data (audio and parameter settings) between two ears at a very fast rate that would enable two hearing aids to synchronize their settings and coordinate for a better feedback identification and noise reduction etc.. Bluetooth connectivity is a recent innovation communicating hearing instruments to audio sources such as, MP3, television, cell phones etc.. Nevertheless, the high power consumption of Bluetooth does not permit its integration into a hearing aid [KCK+10]. The communication between a hearing aid and external audio device is made through a secondary streaming device usually worn around the neck in current hearing aids. Direct communication to audio devices would not only make it possible to transmit the sound directly to the hearing aid, but also gives access for hearing aids to be able to use the computational abilities of those devices, such as mobile phones. It is likely that majority of hearing aids will have wireless receivers embedded in them in future with advances in digital chip technology that will allow for smaller and lower in power digital wireless chips.

Although the auditory perception science and psychoacoustics of hearing impairment are mature fields, the research carried out in these fields have not fully contributed to hearing aid design yet. However, this is a growing area and some auditory models are already being applied in hearing aids within the limitations of computational power of hearing aid DSPs. Loudness models, that would allow HI users to perceive the loudness of a sound same as a NH person by adjusting the frequency-dependent gain, for instance, have been built in hearing aids [Moo10, Dau09]. Auditory scene analysis (ASA) which is defined by Bregman in [Bre94a] as "the organization of sound scenes according to their inferred sources" is the ability to separate individual auditory components from multiple sounds sources. ASA has a growing attention in hearing aid technology [Neu08] which may have many applications that would enhance a HI person's listening in complex environments and is a promising hearing science field for future hearing aids.

The hearing aid research has mostly considered bottom-up (signal-based) approaches which are based on sensory and perceptual processing abilities. However, human auditory perception has also top-down (knowledge-based) processes involving the cognitive system. For instance, results in [BCTW12] showed that signal to noise ratio necessary for correct word recognition varied inversely with the probability of that word occurring in the sentence context. There have been studies that reveal the relation of cognitive abilities (such as memory, attention etc.) to hearing ability, and it has been shown that there is a high correlation between cognitive performance and speech recognition in noise [Lun03]. Considering the fact that each person differs in cognitive abilities and as a result or independently may have different preferences for hearing aid settings, it is inevitable that future hearing aids will be more personalized. Today's hearing aids offer automatic or user controlled adjustments of its settings (such as volume adjustment, classification of the environment etc.), yet, 'learning ability' will be added making them 'intelligent'. The intelligent hearing aids may learn the optimal settings for an individual for different hearing scenarios. All the considerations mentioned in this paragraph, can be addressed in a bigger research field that is 'cognitive science'. We believe that cognitive science will be one of the pioneer research fields for future hearing aid designs, that we call 'cognizant hearing aids'.

Cognitive science is the interdisciplinary study of mind and intelligence that brings together disciplines including Artificial Intelligence, Cognitive Psychology, Neuroscience, Linguistics, Anthropology and Philosophy [Bod06, VTR92]. Mind and knowledge have always been studied in the field of Philosophy since ancient Greek times. The philosophy of mind offers theses about the nature of mind and these theses typically do not result from empirical investigation, yet, subsequently figure in actual empirical investigation in cognitive science [Bec88]. Cognitive anthropology is the study of the relation between human culture and human mind [BHM10]. Color categories, for example, depend on culture-specific cognitive processes [VTR92]. English contains terms for blue and green, while, Tarahumara (a language of northern Mexico) contains a single term meaning blue or green. Linguistics, the scientific study of human language, has become part of cognitive science when the researchers took it as a system of knowledge stored in the memory of a speaker [Jac07]. Cognitive Psychology, Neuroscience and Artificial Intelligence are other three core fields in cognitive science and we will study their role especially in cognitive hearing science in the following paragraphs more in detail as they also cover the research studies included in this thesis.

Cognitive psychology is a sub-discipline of psychology that studies how people think, investigating internal mental processes. The ground-breaking work that covered cognitive and auditory research is thought to be the study performed in the 1950s by Broadbent, Cherry and others in Cambridge [Che53, Bro58], who worked on dichotic listening and selective attention coining the term 'cocktail party phenomena' (one's ability to focus on a particular auditory stimulus while filtering out others) [ALLKPF09]. Listening and language processing in complex environments exploring the interaction of auditory and cognitive factors such as attention, memory (short-term, long-term, semantic), emotion, decision making, perception etc., has often been studied in this field [Bro58, DA07, DCB08, CCB01, ZKK$^+$09, Ake08, SJ01, DFSHB04]. There is accumulated evidence that cognitive functions play a role in listening performance and in the benefits obtained from hearing aids [Ake08, Lun03]. In [CCB01], for instance, they observe that in a cocktail party problem, subjects who detect their name in the irrelevant message have relatively low working-memory capacities suggesting that they have difficulty filtering out distracting information. However, exactly what cognitive factors are important and

how they interact with different hearing-aid settings is currently under investigation [RFST+08].

Neuroscience is the study of the nervous system, that addresses and answers fundamental questions about how the brain and mind work. Although cellular and molecular neuroscience, the study of neurons at a cellular and molecular level, is one of the branches that comes to mind first, there are many other major branches of neuroscience that contributes to cognitive science. Some of these branches are neuroinformatics, the study of application of computational models to neuroscience data, neuroimaging, study on imaging the structure and function of the brain, cognitive neuroscience, study of neural substrates underlying cognition, affective neuroscience, the study of the neural processes involved in emotion and cultural neuroscience, the study of how cultural experience shapes the brain. Computational and theoretical models of cognitive functions such as working memory [OF06], cultural learning [LDI09], auditory and visual attention [KPLL05a, IK01] and learning and decision making [BSW89], have often been studied. Imaging the brain has opened a distinct level of analysis, that could integrate cellular and behavioral models of cognitive science [Pos11]. Recent works have shown that it is possible using electroencephalography (EEG) to recognize emotions [ACS09, PH10b], estimate the direction of auditory attention [EKY+11] and estimate the attended stimulus (speech or music) [LPX+10]. It is apparent that neuroscience studies have a lot to offer for future cognizant hearing aids and could have a major role with wearable EEG systems [PMKT11] communicating hearing aids.

Artificial Intelligence (AI) is the study of machine intelligence and its scientific goal is to understand principles that make intelligent behavior possible in artificial systems, by analyzing natural and artificial agents, formulating hypothesis about construction of intelligent agents and designing and experimenting with computational systems that perform tasks requiring intelligence such as speech communication, vision, decision making, affective computing etc. [PM10]. AI can be applied in a wide range of fields including medical diagnosis, robotics, transportation, telecommunication, music etc.. Current hearing aids already have some automatic features such as turning on and off the directionality, noise reduction, classifying the environment (in, out, home etc.), and ad-

justable settings (volume etc.), yet, learning will be added that would satisfy a user's individual needs more under different situations (environment, mood etc.) [BGH10, Edw07]. In [ZKK+09], they show that partly incorrect or delayed subtitles that are automatically generated by an automatic speech recognition (ASR) system improve speech comprehension of hearing impaired listeners regardless of age and hearing loss. In [DFSHB04], they show that children with hearing impairments show delays in reading minds, recognizing emotions. Therefore we believe that hearing aids can utilize assistive tools such as an ASR system, or an affect recognition system. It is not possible to cover all of the current or future impact of AI studies on hearing aids in this thesis and we believe that it is up to one's imagination to estimate how much AI would affect hearing aids, however, we believe that this effect will be significant.

This thesis will approach the 'cognizant hearing aids' subject in a broad sense with issues related to all cognitive science disciplines but mostly to cognitive psychology, artificial intelligence and neuroscience. It will focus on three specific subjects under these areas, that are automatic speech recognition, top-down auditory attention modeling and automatic emotion recognition from speech.

The thesis is presented in the following structure:

1. **Chapter 2** presents a new approach for robust speaker independent ASR using binary masks, obtained by applying ASA, as feature vectors. The method is evaluated on the isolated digit database, TIDIGIT in three noisy environments (car, bottle and cafe noise). The results are compared with one of the state of the art method using mel frequency cepstral coefficients (MFCC). This chapter includes the work in Appendix A.

2. **Chapter 3** presents a top-down attention model, in which attention is modeled as a sequential decision making process in an environment with features missing completely at random, but there is the possibility to gather additional features among the ones that are missing. We also investigate the role of top-down task driven attention in the cocktail party problem and perform behavioral experiments inspired by Cherry's classic experiments carried out

almost sixty years ago. The participants listen to a mono signal of two different narratives without any given task or with the given task that is to follow just one story, and report words they heard. This chapter includes the work in Appendices B, C, D and E.

3. **Chapter 4** presents a dimensional speech affect recognition model that combines acoustic and semantic features. We design our own corpus that consists of short movie clips just with audio and text (subtitles), rated by human subjects in arousal and valence (A-V) dimensions. This chapter includes the work in Appendix F

4. **Chapter 5** concludes the thesis with a summary.

Chapter 2

Automatic Speech Recognition Using Binary Masks

This chapter describes a new approach for a robust speaker independent ASR system using features from a binary mask. Sections 2.1 and 2.2 state background information, motivation and related work in the literature. Section 2.3 describes the method used for the design including theoretical background. Section 2.4 gives the experimental evaluation results using the isolated digit database, TIDIGIT in three noisy environments (car, bottle and cafe noise) and compares the results obtained with the state of the art ASR feature, known as MFCC. Finally, section 2.5 summarizes the chapter.

2.1 Background

In image pattern recognition tasks, the target patterns vary widely, such as face, shape, object, lightning, etc. In speech pattern recognition, even tough the target pattern variability is less, there are many target patterns to be sought such as speaker identity, the language, the emotional state of the speaker and most often the text translation of what is being said; the research area of which is called automatic speech recognition (ASR).

Some ASR systems focus on a limited number of speakers (sometimes just one speaker), that are called '*speaker-dependent*' systems. In '*speaker-independent*' systems instead, the ASR system does not assume any number of speakers, but the performance of the system is tested independent of who is talking. Because of between-speaker variabilities, a well trained speaker-dependent may outperform a speaker-independent system, yet in many applications a speaker-independent ASR system may be desirable where training for a specific person is not possible or not feasible. Please check [HL93] for a more detailed review of speaker-dependent and speaker-independent ASR systems.

Main difficulty in ASR systems is to handle background and channel noise. Many speech enhancement techniques exist that are successfully applied to noisy speech signals to extract noise and speech signals. Wiener filtering method [TE09], in which the noisy signal is passed through a Finite Impulse Response (FIR) filter whose coefficients are estimated by minimizing the Mean Square Error (MSE) between the clean signal and its estimation to restore the desired signal; spectral subtraction methods [Bol79], in which power or magnitude spectrum of a noise signal is estimated during speech inactive periods and subtracted from the spectrum of the total signal (speech and noise) resulting in the spectrum of the speech; and non-negative matrix factorization [KLPB10a], in which a dictionary of a speech or noise signal is obtained and used to extract the speech signal from the noisy signal; are some of the speech enhancement methods used often.

While automatic sound separation is a challenge, the auditory system has a great performance of sound separation. It is claimed that the auditory system does processing according to an analysis-synthesis strategy

called the auditory scene analysis (ASA) [Bre94b]. Bregman claims that in ASA, the acoustic mixture is decomposed into a collection of segments and following it, these segments are grouped into streams. Inspired by this two stage approach, computational auditory scene analysis (CASA) [WB06] has a sound separation approach of two stages that are called segmentation and grouping. Although, the goal of a perfect CASA system would be to separate all sound sources, this would be beyond human sound separation capacity which is four [Cow01]. Humans can not segregate more than four sound sources simultaneously, and the capacity could be lessened to just one in a cocktail party problem. Ideal binary mask (IBM) is a binary mask in which ones represent time-frequency (T-F) regions in which the energy of the target signal is greater than the energy of the noise (interfering) signal for a *local criteria (LC)* and zeros represent otherwise and Wang in [Wan05b] claims that IBMs are the computational goal of ASA. There are studies, in which the use of IBMs are shown to improve speech intelligibility [WKP$^+$08, WKP$^+$09, KBP$^+$09b]. Then, the attempt to find the answer to whether IBMs can be used as a preprocessor to ASR is quite intuitive.

2.2 Related Work

ASR studies date back to early 50s [O'S08], Bell Labs made a demonstration of a small vocabulary recognition of digits over the telephone in 1952. However, first practical applications that were aimed at recognizing isolated words appeared during 60s. Linear predictive coding (LPC) became the dominant method used during 70s [DPH00]. Even though LPC is still used in ASR, it left its place to Mel-frequency cepstral coefficients (MFCC) during 80s [DM80]. During 80s, a new method emerged that is so called dynamic time warping, that is a method based on a similarity measure between a test and a reference pattern warped (stretched or compressed) nonlinearly and this method was widely used for isolated word recognition [MRR80]. Template methods such as DTW lead to high computational burden, so, they were replaced by statistical models during 90s one of which is the hidden Markov models (HMM) [Rab89]. An HMM is a statistical model that assumes that the system is a Markov process with unknown or i.e. hidden parameters to be determined using observ-

able parameters; and it has been successfully applied for both isolated word and continuous speech recognition areas since 90s [DASW94, Lee90] and is still a widely used method [Che12, SC11, GY08]. Many other methods such as artificial neural networks (ANN) [SCG+10, DS10] and support vector machines (SVM) [SUMIGA+07] have also been successfully applied to ASR systems, yet, use of HMMs is still the most popular approach.

There are open-source and commercial ASR toolkits, such as Sphinx [HAH+92] by University of Carnegie Mellon, HTK [You93] by University of Cambridge and RWTH Aachen University speech recognition toolkit [RGH+09], which made it easier for researchers to build high-quality speech recognition systems [BAS09, AAZ+10]. However, although significant establishments made in ASR for the last 60 years, the performance of the state of the art work is far from human performance; human listeners are much better at dealing with accents, noisy environments, differences in speaking style, speaking rate, etc. [Sch07]. There is still much room for improvement of ASR systems.

2.3 Method

2.3.1 Overview

The modeling framework is summarized in Figure 2.1. We have 11 digit samples pronounced by different male and female speakers as the target speech signal (pure, or with added noise (cafetaria, car, etc.)). If the target sound is noisy, we use the non-negative matrix factorization (NMF) method to extract the target from the mixture. We then use speech shaped noise (SSN) matching the long term spectrum of a large collection of speakers as the reference interference signal to obtain the masks (the details of how we obtain the mask are explained in the following sections). The obtained masks are allocated as train and test sets. As a recognition engine, a discrete hidden Markov model is used, thus a vector quantization model (K-means) is used resulting in a codebook representing all observed outputs for masks of the training set. We train one model for each digit, the test set that is mapped to the obtained

codebook is tested with each model and the test sample is assigned to the one with the highest likelihood calculated.

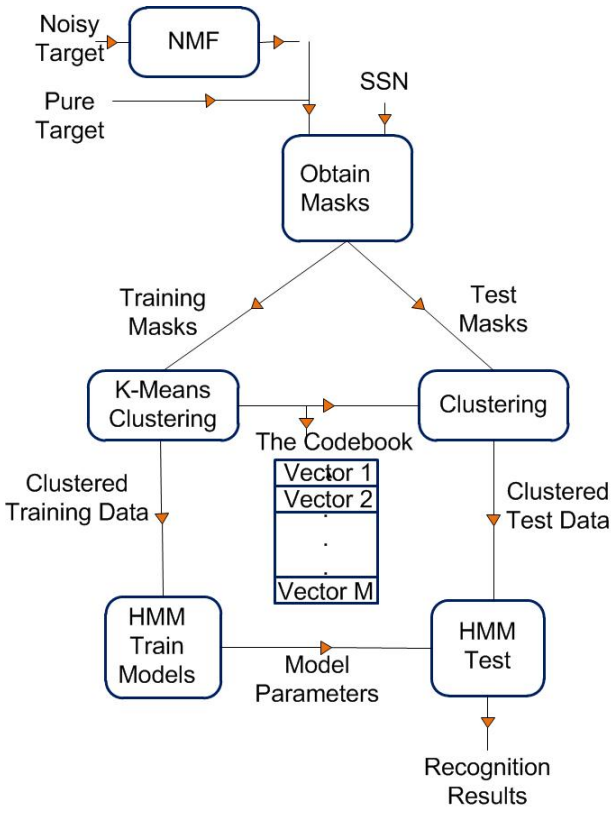


Figure 2.1: The overview of the model used.

2.3.2 Ideal Binary Masks and Target Binary Masks

Ideal binary masks (IBMs), which are denoted as the goal of CASA, are binary representations of a sound signal in time-frequency (T-F) space. The idea of an IBM is to keep the T-F regions of a target sound signal that are stronger than the interference (noise) signal for a value of local criteria (LC) and assign one to those regions, and discard the other regions and assign zero to those. Therefore, if $T(t, f)$ and $N(t, f)$ represent the target

and noise T-F regions, then an IBM is defined as

$$IBM(t,f) = \begin{cases} 1, & \text{if } T(t,f) - N(t,f) > LC \\ 0, & \text{otherwise} \end{cases} \tag{2.1}$$

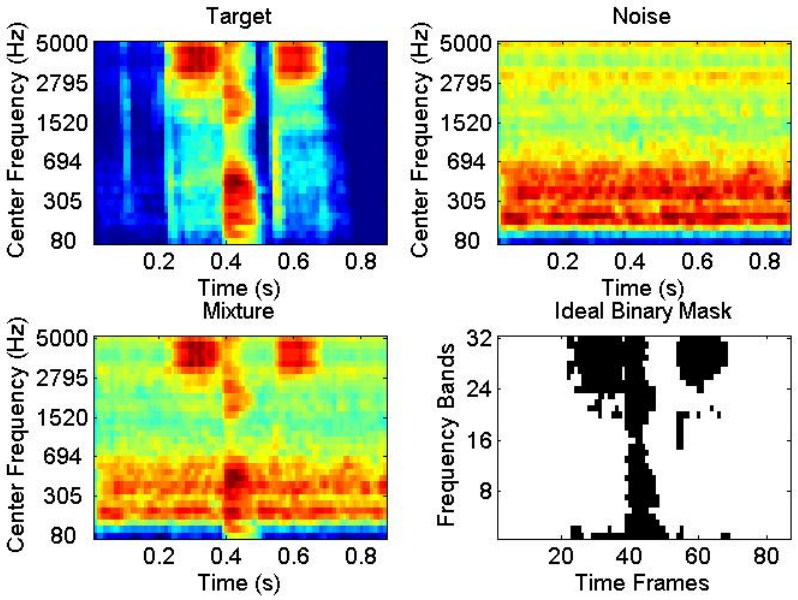


Figure 2.2: The T-F spectrograms of the target signal (digit 'six' pronounced), noise signal (SSN), and the mixture signal (target+noise) and the resultant IBM.

Figure 2.2 shows an example of an IBM for digit 'six'. LC value is crucial, very high LC values could lead to all-zeros mask and very low LC values could lead to all ones mask. In Figure 2.3 (taken from [KBP+09b]), we can see the speech intelligibility scores as a result of study in [KBP+09b]. In that study they make the participants listen to IBM-gated speech signals (they multiply the T-F spectrum of the noisy speech signal with its IBM, and resynthesize it) As seen in the figure, the resultant scores are bell-shaped for different LC values. The intelligibility is high for some LC values and decrease with either increasing or decreasing that value since the mask is getting close to all-zeros or all-ones mask. Therefore the LC

value should be optimized for an application also depending on the signal to noise ratio (SNR) (the higher the SNR the higher the LC should be to obtain the same mask). In [KBP+09b], they define a relative criteria $RC = LC - SNR$, and they reported high human speech intelligibility (over 95%) for the RC range of [-17 dB, 5 dB]. In that figure, we also observe that the speech intelligibility score is very high even for SNR of -60dB with the right LC values leading to the conclusion that the IBM pattern itself might be recognized by humans and we investigate in this work if ASR could use this pattern.

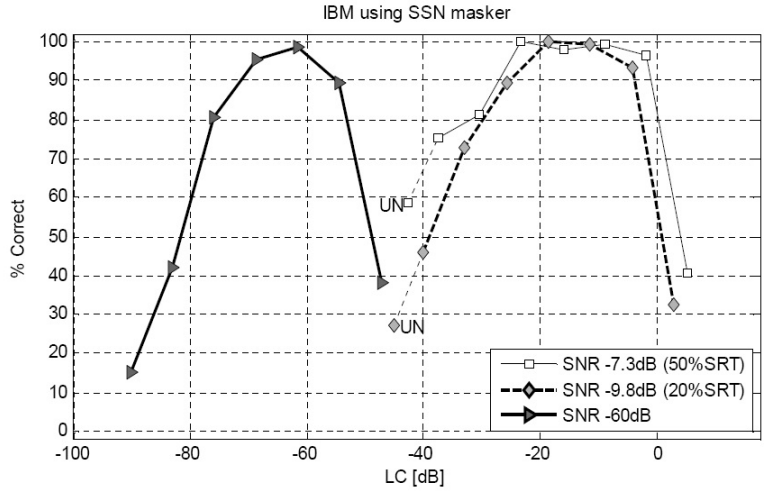


Figure 2.3: Intelligibilty score versus LC for IBM with SSN from the study in [KBP+09b].

Calculating an IBM requires that the target and the noise are available separately, which is a major drawback. An alternative mask definition has been made which uses a reference SSN signal matching the long term spectrum of the target speaker as the noise signal whatever the real noise type is and it is called target binary mask(TBM) [KBP+09b]. TBMs have also been showed to lead to high speech intelligibility scores [KBP+09b]. The definition of a TBM is very similar to an IBM shown as below:

$$TBM(t,f) = \begin{cases} 1, & \text{if } T(t,f) - SSN(t,f) > LC \\ 0, & \text{otherwise} \end{cases} \tag{2.2}$$

where, only prior target signal information is compulsory. Figure 2.4 shows IBM and TBM of a digit 'three'. It is observed that there is no crucial difference between the resultant masks and, using TBMs, we could avoid to train HMMs for each noise type (Note that IBM is identical to TBM when the noise type is SSN) in our work. More figures of IBMs can be found in Appendix A for different digit samples and different speakers.

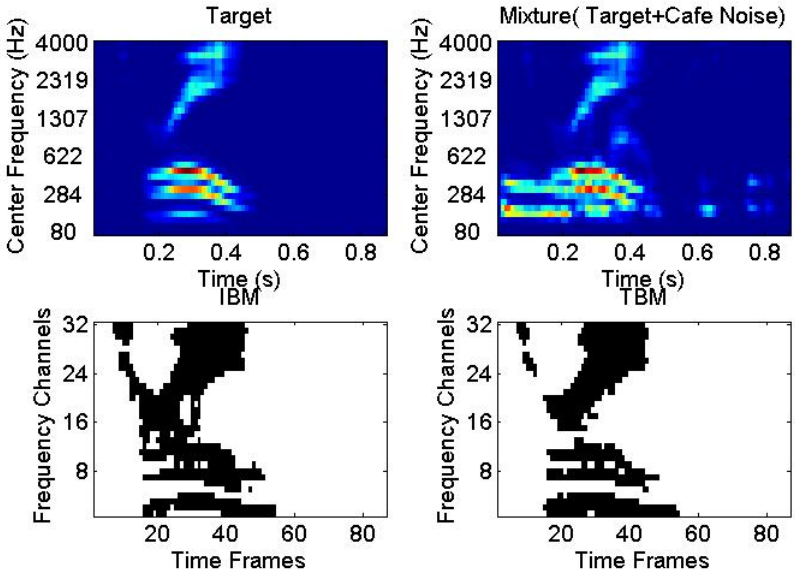


Figure 2.4: The T-F spectrograms of the target signal (digit 'three' pronounced), the mixture signal (target + cafe noise) and the resultant IBM and TBM.

2.3.3 ASR Model

We use a discrete Hidden Markov Model (HMM) as the recognition engine [Rab89]. Figure 2.5 gives an example of an HMM with N states and M observed outputs. An HMM can be an ergodic model in which a state can be reached via any other state, or a left-to-right model in which the states proceed from left to right which is more appropriate for source signals whose properties change over time such as speech. An HMM is characterized by :

1. N, the number of states in the model
2. M, the number of observation symbols (outputs)
3. A=a(i,j) where $1 \leq i, j \leq N$, the state transition probability distribution
4. B=b(k) where $1 \leq k \leq M$, the observation symbol probability distribution
5. π where $\pi_i = P(q_1 = S_i)$, the initial state distribution (q_1 is the starting state)

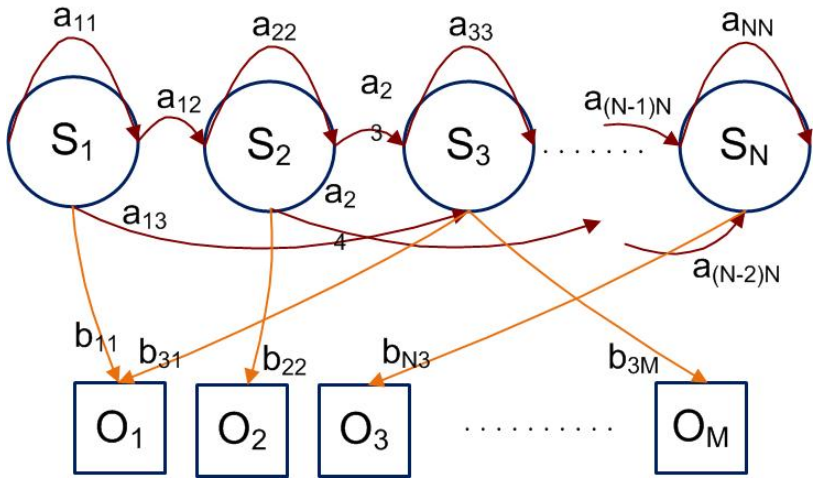


Figure 2.5: An illustration of a hidden Markov model.

In our project, the source signal is the speech signal, and the observed output is the binary mask. The discrete observation symbols are obtained using the K-means algorithm as a vector quantization method. The K-means algorithm is well known for unsupervised data classification, especially for large data sets and is very simple to apply [Moo01].

The data to be clustered are binary vectors obtained from the masks. The columns of the masks are concatenated as in Figure 2.6 and the number of columns to be concatenated is a parameter to be optimized for the optimal solution. Then, the vectors are clustered into M (the number of observation symbols of an HMM) number of classes.

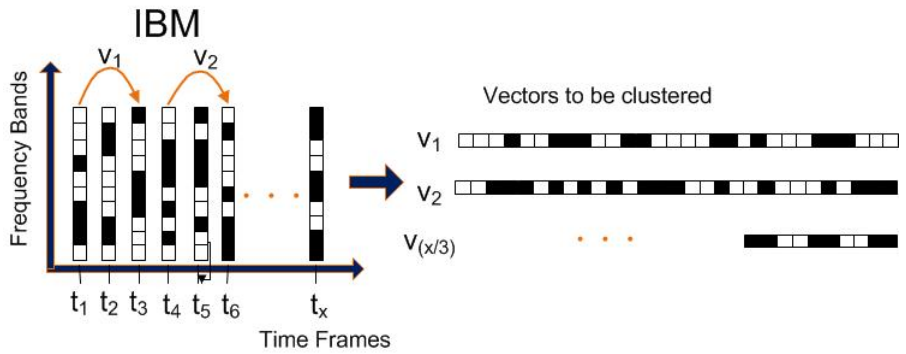


Figure 2.6: Illustration of acquisition of observed outputs to be clustered by K-means using IBMs

2.3.4 Estimation of TBMs

Non-negative sparse coding (NNSC) [Hoy02], a combination of sparse coding and non-negative matrix factorization, is used to extract the target speech information from a noisy speech signal. This method was proven to be successful for wind noise reduction in [SLH07].

The principle in NNSC is to factorize the non-negative signal, X into a dictionary W and a code H:

$$X \approx WH. \tag{2.3}$$

The columns of the dictionary matrix constitute a source specific basis and the sparse code matrix can be considered to have the weights for each of the basis vectors constituting the signal X, that is the T-F spectrograms in our work. Enforcing the code to be sparse results in a solution where only a few of the dictionary elements are active simultaneously, forcing the dictionary elements to be more source specific. To compute this factorization, the algorithm explained in [EK04] and used in [SLH07]. W and H are initialized randomly, and updated according

to the equations below until convergence:

$$H \longleftarrow H \times \frac{W^T.X}{W^T.W.H + \lambda}, \tag{2.4}$$

$$W \longleftarrow W \times \frac{X.H^T + W \times (1.(W.H.H^T \times W))}{W.H.H^T + W \times (1.(X.H^T \times W)))}, \tag{2.5}$$

where, $(.)$ indicate direct multiplication, while $(\times)$ and $(\_)$ indicate point wise multiplication and division. 1 is a square matrix of ones of suitable size.

When the speech signal is noisy, and if the noise signal is assumed to be additive, then

$$X = X_s + X_n \approx [W_s W_n] \begin{bmatrix} H_s \\ H_n \end{bmatrix}, \tag{2.6}$$

where X_s and X_n denote the speech and noise. We compute the noise dictionary W_n using noise signal samples we have, and using equations 2.4 and 2.5. We keep this precomputed W_n fixed and learn speech X_s using the following iterative algorithm,

$$H_s \longleftarrow H_s \times \frac{W_s^T.X}{W_s^T.W.H + l_s}, \tag{2.7}$$

$$H_n \longleftarrow H_n \times \frac{W_n^T.X}{W_n^T.W.H + l_n}, \tag{2.8}$$

$$W_s \longleftarrow W_s \times \frac{X.H_s^T + W_s \times (1.(W.H.H_s^T \times W_s))}{W.H.H_s^T + W_s \times (1.(X.H_s^T \times W_s)))}, \tag{2.9}$$

The clean speech is estimated as

$$X_s = W_s H_s. \tag{2.10}$$

Figure 2.7 shows the T-F spectrograms of a pure, noisy, estimated target signal (digit 'six') and of the noise signal (bottle noise that is the sound of many bottles chinking on a production line). Finally, TBM is estimated using the estimated target speech signal X_s and the reference SSN signal as in equation 2.2, assuming that we have pure target signal spectrogram after applying NNSC. As mentioned before, LC and SNR values are crucial, to have the maximum performance using binary masks. We define a new SNR in T-F domain and call it SNR_{TFD} so that we could avoid to going back to the time domain that could result in a waste of time and computational power. SNR_{TFD} is calculated as the ratio between the sum of all T-F bins of the target signal to the sum of all T-F bins of the noise signal.

2.4 Experimental Evaluation

The dataset we use in order to evaluate our work is TIDIGITs [Leo84]. We use 11 digit samples (oh, zero, one, two, three, four, five, six, seven, eight and nine) uttered by 74 male and 100 female speakers. Each digit sample is uttered twice by each speaker making a dataset of 348 samples. We divide the dataset into 174 train, 87 verification and 87 test samples. HMM toolbox by Kevin Murphy for Matlab [Mur98] and NMF toolbox by Technical University of Denmark [IMM06] was used. The sound signals were resampled at 8kHz and the T-F representations were obtained using gammatone filters, an auditory filter simulating the response of a basilar membrane [PHA92], with 32 frequency channels equally distributed on the ERB scale within the range of [80 Hz, 4000 Hz]. The output from each filterbank channel was divided into 20 ms frames with 10 ms overlap. The noise signals used are SSN, as well as car, cafe and bottle (the sound of many bottles chinking on a production line) sounds that are parts of the The DRCD Sound Effects Library [Rad].

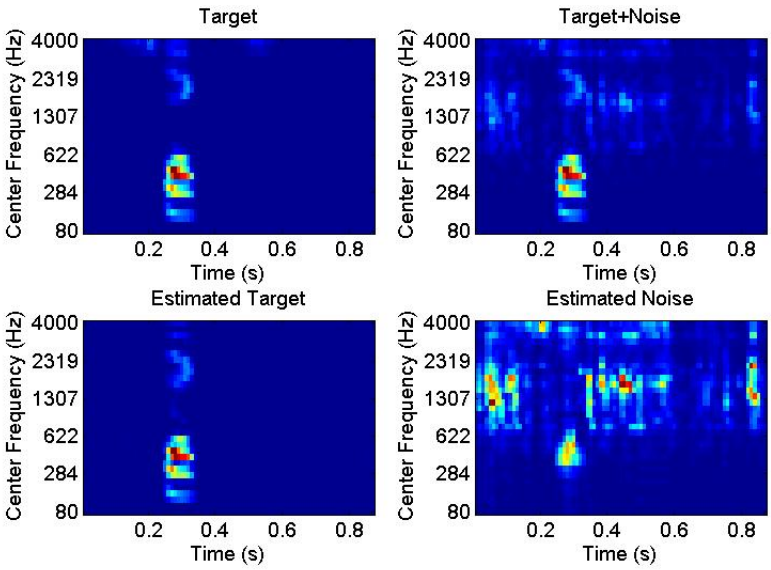


Figure 2.7: T-F spectrograms of a pure, noisy, estimated target signal (digit 'six') and of the noise signal (bottle noise that is the sound of many bottles chinking on a production line), with SNR of 0dB, noise dictionary W_n of 32

The verification set was used to optimize the parameters in the model we used for a higher recognition rate and less computational burden. The optimized parameters are 10 for the number of states in the HMMs trained, 256 for the codebook size of K-means, 3 for the number of mask columns (see Figure 2.6) to be concatenated to obtain the codebook after K-means, 20 ms with 10 ms overlaps for the window length (the length of time slots of a mask), 32 for the number of frequency channels in a mask and 32 for the noise dictionary size W_n. The LC value is very crucial as explained before and depends highly on the SNR value. We take the results of the work in [KBP$^+$09b] as in Figure 2.3 as reference to determine the SNR and LC values that resulted in high speech intelligibility. The SNR and LC values are kept at 0 dB and -8 dB respectively during the optimization process. The recognition results are evaluated for SNR values in the range of [0dB, 20dB] and for LC values in the range of [-4dB, 2dB]. The method is compared with a standard approach using 20 static MFCC features. MFCC features are applied in the same way as binary mask features.

Figure 2.8 shows the recognition rates using IBM or MFCC features as explained before. The HMMS were trained with clean target signals, while tested with added noise at differeny SNR values. It can be seen that using TBM features yields more noise-robust recognition rates than using just MFCC features. It should be noted that we did not apply any improvement methods for MFCC features such as using its dynamical properties that would make it more robust to added noise [YSL07]. However, we did not use any dynamical properties of TBMs either. Although, we make a rough comparison here without using recent improvements for MFCC features, we believe that the comparison is fair since we think there is much room also to improve the use of TBM features as well. We also believe the importance of searching for new routes in ASR.

Figure 2.9 shows the recognition rates using TBMs (or IBMs with SSN), for different SNR_{TFD} values and a fixed LC value of 0dB. We observe that the recognition performance is at its maximum for an RC value of -6dB (LC=0dB, SNR=6dB). Therefore, for a pure target signal, if we keep the RC value around -6dB, we could obtain high performance. However, in real life applications SNR value is unknown, thus it is not trivial to determine the optimal LC value. Assuming we have the noise

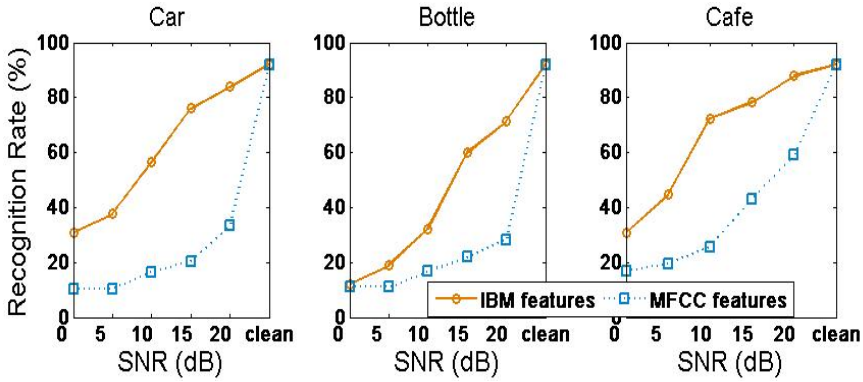


Figure 2.8: The recognition rates for TBMs (IBMs with SSN) and MFCC features for car, bottle and cafe noise types, with optimized LC value for the best performance

characteristics, we use NNSC to solve this problem. As explained before, we assume that we have the pure target signal spectrogram after applying NNSC. Figure 2.10 shows the recognition rates before and after applying NNSC using TBMs for three different noise types (HMMs are trained with clean data, the test is carried out with noisy data). Before NNSC, using different LC values constituting an RC range of [-4dB, 2dB] that was shown to yield high recognition results testing with clean data (please see Figure 2.9), results in very scattered rates for a fixed SNR value (e.g. 'bottle' noise at 20dB of SNR, rates between 20% to 65% for different LC values). After NNSC instead, we have more stable results in terms of use of different LC values which makes the choice of LC less challenging. Moreover, using NNSC as a preprocessor to ASR, also yields higher recognition performance especially for noisier signals. Finally, we obtain 60%to 70%, 16% to 73% and 40% to 70% recognition rates for SNR values between 0 dB and 20 dB for car, bottle and cafe noises respectively, which are comparable to the state-of-the-art results [YSL07, GP06].

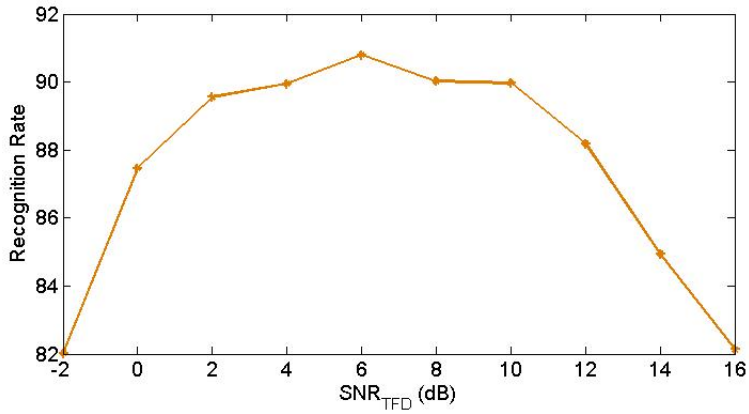


Figure 2.9: The recognition rates with TBMs for LC=0 dB, SNR_{TFD}=[-2 dB, 16 dB], W_n=64, W_s=128

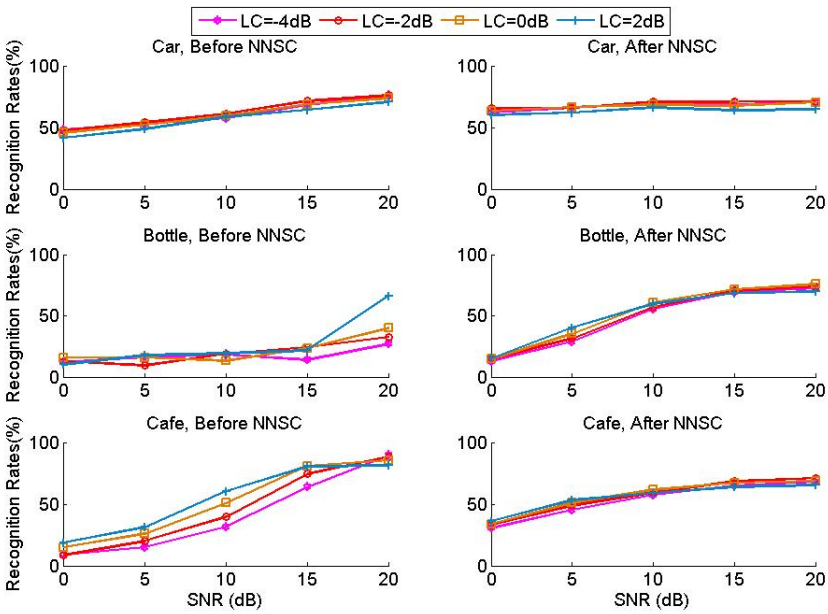


Figure 2.10: The recognition rates before and after NNSC, for SNR_{TFD} of 0dB

2.5 Summary

We investigated a new feature extraction method for ASR using ideal and target binary masks. We built a speaker-independent isolated digit recognition system that used discrete HMMs as the recognition engine. K-means algorithm was used for vector quantization method. Isolated digit samples, pronounced by 87 different male and female speakers, from the TIDIGIT database were used, and split into a train, verification and test sets. The system was also tested with digit samples with added noise that are of type cafe, bottle and car from the DRCD Sound Effects Library. We used non-negative sparse coding method to extract target signals from noisy signals.

The maximum recognition rate achieved with clean speech signals is 92%. This work also showed that the success of this recognition method was not limited to the ideal case where the target and the noise signals were available separately. We obtained over 60% of recognition rates for different noise types, even for SNR ratios as low as 0dB, using estimated TBMs. We concluded that using binary information from the masks directly can lead to noise-robust ASR systems and there is room for further improvement.

Chapter 3

Top-Down Auditory Attention Modeling

This chapter describes the computational model for top-down attention we designed and behavioral experiments we carried out with human subjects in order to explore top-down attention in human hearing. Sections 3.1 and 3.2 state background information, motivation and related work in the literature. Section 3.3 explains the methods we used for both computational and experimental parts. Section 3.4 gives the experimental evaluation results using the computational model with some well-known classification problems (the UCI repository [FA10]) and the results obtained by analyzing the behavioral experiments carried out. Section 3.5 finally summarizes the chapter.

3.1 Background

The word 'attention' has a Latin root 'attentus' which means to 'heed'. For a living creature, one of the key points of survival is to be able

to attend to the most relevant stimuli in a complex environment. The attention mechanism of many species, especially of humans, attracted scientists and philosophers for hundreds of years. The first extended treatment of attention may be due to Malebranche [Mal64] who claimed that '*attention is necessary for conserving evidence in our knowledge*'. Since then, attention has been studied broadly by many researchers, yet, a universal definition of it on which everyone agrees is still not made. Pashler [Pas99] claims that '*no one knows what attention is, and that there may even not be an 'it' there to be known about (although of course there might be)*'. [1]

Units of attention have been discussed for a long time and three main theories exist. Some studies suggest that attention is based on features (feature based attention) [TG80] attending particular features such as color, direction, pitch, etc.; some suggest that it is based on location (location based attention, spatial attention) attending selectively to information in a particular physical location [Pos80, DAM11]; and others suggest it is based on objects (object based attention) [BZM98] attending selectively to an individual object in the scene. There are also studies on these attention mechanism working together [BZM98, ACMR00].

Attention is about selecting the most salient stimuli in a complex environment while ignoring the others as mentioned before. There are two main conceptual principles in this selection process, that are *bottom-up* and *top-down*. According to bottom-up perspective, attention is driven mainly by the characteristics of the target stimulus and its sensory context independent of an aim, task or previous knowledge, like attending to red point in a green background, to a scream sound with a relatively silent background, etc. [SGB01]. Meanwhile, in top-down perspective, attention is driven by previous knowledge, aim, task, emotions or desires and the subject usually develops expectations, like attending to the music sound if the task is to find out which instrument is being played , to round objects if the task is to find the oranges in a scene, etc. [Ung00, SGB01].

Human attention is considered to develop as a consequence of evolution to enable humans to survive by dealing with massive amount of data and select the salient ones ignoring the rest. Dealing with massive amount

[1] For a more detailed review of attention please see [IRT05]

of data is also a problem in modern technical systems, such as computer vision systems, autonomous robots, etc. and it is crucial to reduce data, or assign saliency to each data unit. Therefore, human attention mechanism has taken a lot of interest from researchers for years, and computational models have been developed of both visual and auditory attention considering both top-down and bottom-up perspectives. Computational modeling of attention has started with the Feature Integration Theory of Treisman and Gelade [TG80], in which they proposed that only simple visual features are computed in a parallel manner over the visual field. The interest in this field increased over the years, one example is the work by Itti [Itt04], in which a biologically-motivated algorithm is designed to select visually-salient regions of interest in video streams, and another is example is the work by Kalinli and Narayanan [KN07], in which a saliency based auditory attention model is designed for syllable detection in speech [2].

3.2 Related Work

In auditory attention research field, the well known *cocktail party phenomenon*, the human ability to attend to a single talker among various conversations and noise, has been studied often and experiments were designed with human subjects [Che53, Mor59, WC95]. Around sixty years ago, Cherry [Che53] carried out experiments studying this phenomenon, the results of which are still of importance till this day. In the first set of experiments, the participants listened to a monaural mixture of two narratives spoken by the same speaker and repeated what was being said at the same time (*shadowing effect*). In the second set, they listened to a stereo mixture of two narratives (one narrative in one ear, *dichotic listening*). Cherry found that the subjects who shadowed one message presented in one ear, ignored the second message. The subjects were just able to detect some physical features such as gender of the voice, speech or noise, etc. In 1959 Moray [Mor59], made experiments in exploring attention in dichotic listening, using the shadowing method. Moray also observed that shadowing one message, the subjects were not able to report any of the content of the rejected message, even if the message contained

[2] For a more detailed review of attention modeling please see [Mar12]

just a list of simple words repeated. Meanwhile, he reported that the only stimulus that broke that barrier was the subject's own name pronounced.

Theories have been built to explain the results of human auditory attention experiments. Broadbent [Bro58] explained the results of Cherry in 1958, with an *early filter* theory, that is based on the existence of a filter operating before any semantic interpretation of the stimuli. Meanwhile, Deutsch and Deutsch in 1963 [DD63], built the *late filter* theory claiming that all the stimuli were analyzed semantically and the selection took place afterwards. Finally, Treisman in 1960 [Tre60] introduced the *attenuating filter* theory claiming that the information in the ignored message is not eliminated but attenuated and delayed.

The computational modeling of human attention mechanism has been explored in both bottom-up and top-down perspectives. In bottom-up approaches, in which one attends what pops out, the model is built considering a *saliency map*, encoding the stimuli for salience. Koch and Ullman in [KU85], Itti in [IKN98] used this approach for visual attention modeling, while Kayser in [KPLL05b], Duangudom and Anderson in [DA07] used saliency maps for auditory attention modeling. There have been efforts on building top-down attention models as well although bottom-up approach has been investigated more often. In [Fri06], Frintrop designs a model that combines bottom-up cues (strong contrasts, uniqueness, etc.) and top-down cues (goal directed search, like looking for a yellow hat etc.) for object recognition purpose. Coensel and Botteldooren, in [DCB08], designed a top-down attention model in the auditory domain, which is an immature area not studied often, combining again bottom-up, sound-based cues and top-down, experience dependent cues (expected context), in which meaning is attached as well.

3.3 Method

3.3.1 Overview

We designed a computational model for top-down attention and carried out experiments with human subjects to explore top-down attention in

human hearing [Mar12].

We model top-down attention as a sequential measurement problem (as a classification problem) in which information is missing (not complete) and active decision process takes place to decide what to measure next. In other words, we simulated an environment in which data is missing, and we searched for what to attend among those missing data. The top-down attention model is implemented using a generic Gaussian discrete mixture model and the model is trained with either complete or missing data, while tested with missing data. The model is tested with either artificially created data or some well-known datasets from UCI machine learning repository [FA10].

We designed some behavioral experiments to understand the influence of top-down attention cues in a cocktail party problem, inspired by Cherry's work [Che53]. We made the participants listen to two narratives as a mono signal with reduced spatial and speaker cues in both ears. Two sets of experiments were carried out. In both sets of experiments, the participants were asked to follow the sound signal and report the words they heard, yet, in one of them they were given a task (asked to follow just one of the stories). We explored the effect of presence of a task by analyzing the relation between the number of times the words heard and the temporal and spectral overlaps of the two sound signals mixed for the corresponding words. Finally, we simulated the existence of a task in our computational model in order to compare with results of the experiments that revealed human behavior.

3.3.2 Computational Model Proposed: Top-down attention as a sequential measurement problem

3.3.2.1 Modeling Framework

We used a Gaussian Discrete mixture model (GDMM), see e.g., [HSK$^+$00]. Define $\boldsymbol{x}$ as the d-dimensional input feature vector and the associated output, $y \in \{1, 2, \cdots, C\}$, of class labels, assuming C mutually exclusive

classes. The joint input/output density is modeled as

$$p(y, \boldsymbol{x}|\boldsymbol{\theta}) = \sum_{k=1}^{K} P(y|k)p(\boldsymbol{x}|k)P(k) \tag{3.1}$$

$$p(\boldsymbol{x}|k) = \frac{1}{\sqrt{|2\pi\boldsymbol{\Sigma}_k|}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_k)^\top \boldsymbol{\Sigma}_k^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_k)\right), \tag{3.2}$$

where K is the number of components, $p(\boldsymbol{x}|k)$ are the component Gaussians mixed with the non-negative proportions $P(k)$, $\sum_{k=1}^{K} P(k) = 1$ and the class-cluster conditionals are denoted by $P(y|k)$. $\boldsymbol{\theta}$ is the vector of all model parameters, i.e., $\boldsymbol{\theta} \equiv \{P(y|k), \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, P(k) : \forall k, y\}$. The joint input/output for each components is assumed to factorize, i.e., $p(y, \boldsymbol{x}|k) = P(y|k)p(\boldsymbol{x}|k)$.

The input density associated with Eq. (3.1) is given by

$$p(\boldsymbol{x}|\boldsymbol{\theta}_u) = \sum_{y=1}^{C} p(y, \boldsymbol{x}) = \sum_{k=1}^{K} p(\boldsymbol{x}|k)P(k),$$

where $\boldsymbol{\theta}_u \equiv \{\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, P(k) : \forall k, y\}$. Assuming a 0/1 loss function the optimal Bayes' classification rule is $\widehat{y} = \max_y P(y|\boldsymbol{x})$ where[3]

$$P(y|\boldsymbol{x}) = \frac{p(y, \boldsymbol{x})}{p(\boldsymbol{x})} = \sum_{k=1}^{K} P(y|k)P(k|\boldsymbol{x}),$$

with $P(k|\boldsymbol{x}) = p(\boldsymbol{x}|k)P(k)/p(\boldsymbol{x})$.

Assume that we are given a data set of independent i.i.d. input-output examples $\mathcal{D} = \{\boldsymbol{x}_n, y_n; n = 1, 2, \cdots, N\}$. The negative log-likelihood is given by

$$L = -\log p(\mathcal{D}|\boldsymbol{\theta}) = -\sum_{n\in\mathcal{D}} \log \sum_{k=1}^{K} P(y_n|k)p(\boldsymbol{x}_n|k)P(k)$$

[3]The dependence on $\boldsymbol{\theta}$ is suppressed for clarity.

We use an iterative modified EM algorithm to estimate the model parameters [HSK+00]:

1. For initialization, the data is modeled as a normal distribution with parameter $\boldsymbol{\mu}_0$ and $\boldsymbol{\Sigma}_0$. Only observed data is used during calculations in the case of missing data. However, $\boldsymbol{\Sigma}_0$ is regularized since the estimated covariance matrix is unbiased and is not guaranteed to be positive-semi definite. The regularization is based on inflating the diagonal elements, similar to the approach presented in [Sch01]. Further details can be found in our work in Appendix B. We draw random samples from the distribution, $\boldsymbol{\mu}_0$ and $\boldsymbol{\Sigma}_0$ and apply the KKZ method [SD07, KJKZ94], that is based on paying attention to the data points that are furthest apart from each other, since those data points are more likely to belong to different clusters. Last, $P(k) = 1/K$ and $P(y|k)$ is randomly initialized.

2. Compute posterior component probability for all $n \in \mathcal{D}$:
$$P(k|y_n, \boldsymbol{x}_n) = \frac{P(y_n|k)p(\boldsymbol{x}_n|k)P(k)}{\sum_k P(y_n|k)p(\boldsymbol{x}_n|k)P(k)}. \tag{3.3}$$

3. For all k, update mean vectors and covariance matrices
$$\boldsymbol{\mu}_k = \frac{\sum_{n\in\mathcal{D}} \boldsymbol{x}_n P(k|y_n, \boldsymbol{x}_n)}{\sum_{n\in\mathcal{D}} P(k|y_n, \boldsymbol{x}_n)}, \qquad \boldsymbol{\Sigma}_k = \frac{\sum_{n\in\mathcal{D}} \boldsymbol{S}_{kn} P(k|y_n, \boldsymbol{x}_n)}{\sum_{n\in\mathcal{D}} P(k|y_n, \boldsymbol{x}_n)}$$
where $\boldsymbol{S}_{kn} = (\boldsymbol{x}_n - \boldsymbol{\mu}_k)(\boldsymbol{x}_n - \boldsymbol{\mu}_k)^\top$.

4. For all k update cluster priors and class cluster posteriors
$$P(k) = \frac{\sum_{n\in\mathcal{D}} P(k|y_n, \boldsymbol{x}_n)}{N}, \qquad P(y|k) = \frac{\sum_{n\in\mathcal{D}} \delta_{y-y_n} P(k|y_n, \boldsymbol{x}_n)}{\sum_{n\in\mathcal{D}} P(k|y_n, \boldsymbol{x}_n)}$$

3.3.2.2 Missing Data Problem

Missing data problem has been faced in a broad range of studies such as speech recognition from degraded (noisy etc.) speech signal, medical

diagnosis using machine learning applications with incomplete surveys and social data analysis with incomplete interviews. Rubin, in [Rub76], regarded missingness as a probabilistic phenomenon as below. Let us define the distribution of missingness as M, complete data as X_{com}, and missing and observed parts of data as X_m and X_o,

- Missing at random (MAR): The probability of missingness may depend on observed data but not on missing data.

$$P(M|X_{com}) = P(M|X_o)$$

- Missing completely at random (MCAR): The probability of missingness may depend on neither observed nor missing data.

$$P(M|X_{com}) = P(M)$$

- Missing not at random (MNAR): The probability of missingness may depend on missing data.

There are three main missing data techniques (MDT) to deal with missing data.

- Deletion: In this method, the missing data is discarded and the analysis is done on the observed data. In *pairwise deletion*, only the missing features are deleted, while in *listwise deletion*, all the samples including one or more missing features are deleted. In the former method, all available information is used, yet, it is difficult to obtain accurate density estimations. In the latter, density estimation is easy since the analysis is based on only complete samples, yet, potentially usable data is discarded which could lead to insufficient data samples.
- Imputation: In this method, a new value is assigned to each missing value. The most used and simplest imputation method is *mean imputation*, in which the missing value of a variable is replaced by the mean of that variable among observed data. This method has a drawback of producing biased estimates. Another well known method is *multiple imputation*, in which multiple values are replaced for a missing point leading to alternative versions of the complete data.

- Model-based: In this method, the analysis can be made directly on the incomplete data without any imputation or deletion. In *Maximum Likelihood* approach, the data is modeled on observed data and the missing ones are estimated using this distribution.

For a more detailed review of missing data problem, please see [SG02] by Schafer and Graham. We compared some of the well known MDTs in our work in [KMHL11] which can also be seen in Appendix B. Using our comparison results, in this work, we decided to use a maximum likelihood method that we call *complete expectation maximization (CEM)*, proposed by Ghahramani and Jordan in [GJ94]. Missing data compensation is carried out within the EM step of the model where also the model components are estimated. The posterior component probability, $p(k|y_n, \boldsymbol{x}_n)$, is again calculated as in Eq. 3.3, but only on observed dimensions. Let ^{o} and ^{m} represent observed and missing parts. Then, for updating the mean vector, $E[\boldsymbol{x}_n^m|\boldsymbol{x}_n^o]$ substitutes missing components of $\boldsymbol{x}_n$, and for updating the covariance matrix, $E[\boldsymbol{x}_n^m \boldsymbol{x}_n^{m^T}|\boldsymbol{x}_n^o]$ is substituted for outer product matrices containing the missing components

$$E[\boldsymbol{x}_n^m|\boldsymbol{x}_n^o] = \mu_k^m + \boldsymbol{\Sigma}_k^{mo}\boldsymbol{\Sigma}_k^{oo^{-1}}(\boldsymbol{x}_n^o - \mu_k^o),$$

$$E[\boldsymbol{x}_n^m \boldsymbol{x}_n^{m^\top}|\boldsymbol{x}_n^o] = \boldsymbol{\Sigma}_k^{mm} - \boldsymbol{\Sigma}_k^{mo}\boldsymbol{\Sigma}_k^{oo^{-1}}\boldsymbol{\Sigma}_k^{mo^T} + E[\boldsymbol{x}_n^m|\boldsymbol{x}_n^o]E[\boldsymbol{x}_n^m|\boldsymbol{x}_n^o]^\top.$$

3.3.2.3 Attention Model

Kappen et al. [KNvM98] considered the sequential measurement process as a two-stage process as we do in this work. As shown before, we represent the signal detection problem by a probability distribution over a set of C classes indexed by the discrete variable $y, (y = 1, ..., C)$. However, we split the input data in observed and un-observed as follows. Initially, the observation $\boldsymbol{x}$ with components $x_i, (i = 1, ..., I)$ is available. Additional measurement z_j is chosen among the set of missing features $\boldsymbol{z}$ with components $z_1, ..., z_J$ in the second step of the process.

Denote the joint probability of the classes and all features observed and missing as $p(y, \boldsymbol{x}, \boldsymbol{z})$. We use Bayesian decision theory and use the posterior distribution $p(y|...)$ to make inferences about y. The condition

depends on the stage in the sequential measurement process. Initially, the information available is $\boldsymbol{x}$, thus the relevant probability is

$$\begin{aligned} p(y|\boldsymbol{x}) &= \int p(y, \boldsymbol{z}|\boldsymbol{x}) d\boldsymbol{z} \\ &= \frac{\int p(y, \boldsymbol{x}, \boldsymbol{z}) d\boldsymbol{z}}{\sum_{y=1}^{C} \int p(y, \boldsymbol{x}, \boldsymbol{z}) d\boldsymbol{z}}. \end{aligned} \tag{3.4}$$

We select an additional feature $\boldsymbol{z}_j$ using the top down attention model, that will result in the distribution

$$\begin{aligned} p(y|\boldsymbol{x}, z_j) &= \sum_{y=1}^{C} \int p(y, \boldsymbol{z}|\boldsymbol{x}) \prod_{i \neq j} dz_i \\ &= \frac{\int p(y, \boldsymbol{x}, \boldsymbol{z}) \prod_{i \neq j} dz_i}{\sum_{y=1}^{C} \int p(y, \boldsymbol{x}, \boldsymbol{z}) \prod_{i \neq j} dz_i} \end{aligned} \tag{3.5}$$

The difference in confusion (entropy) before and after the second measurement, is the information obtained by the additional measurement, and this is the quantity we want to maximize

$$\begin{aligned} \Delta S_j(\boldsymbol{x}, z_j) &= \sum_{y=1}^{C} \log p(y|\boldsymbol{x}, z_j) p(y|\boldsymbol{x}, z_j) \\ &- \sum_{y=1}^{C} \int \log p(y, \boldsymbol{z}|\boldsymbol{x}) p(y, \boldsymbol{z}|\boldsymbol{x}) d\boldsymbol{z} \end{aligned} \tag{3.6}$$

Since we do not know z_j at this step, we average $\Delta S_j(\boldsymbol{x}, z_j)$ with respect to this variable given the information we have, i.e., with respect to the distribution of z_j conditioned on the initial measurement $\boldsymbol{x}$. This procedure provides us with the *expected information gain* of measuring the value of feature j,

$$\begin{aligned} G_j(\boldsymbol{x}) &\equiv \int \Delta S_j(\boldsymbol{x}, z_j) p(z_j|\boldsymbol{x}) dz_j \\ &= \sum_{y=1}^{C} \int \log p(y|\boldsymbol{x}, z_j) p(y, z_j|\boldsymbol{x}) dz_j \\ &\quad - \sum_{y=1}^{C} \int \log p(y, \boldsymbol{z}|\boldsymbol{x}) p(y, \boldsymbol{z}|\boldsymbol{x}) d\boldsymbol{z}. \end{aligned} \tag{3.7}$$

The information gain can be used to rank features in importance [Lin56, KNvM98]. The second term can be neglected in the saliency estimate since it does not depend on j.

Introducing the Gaussian discrete model (Eq. (3.1)) within the expression for the information gain and using $p(k|\boldsymbol{x}) = p(k)p(\boldsymbol{x}|k)/p(\boldsymbol{x})$, we obtain

$$\begin{aligned} G_j(\boldsymbol{x}) &= \sum_{y=1}^{C}\sum_{k=1}^{K} p(y|k)p(k|\boldsymbol{x}) \times \\ &\qquad \int \log\left[p(y,\boldsymbol{x},z_j)\right] p(z_j|\boldsymbol{x},k)dz_j \\ &- \sum_{k=1}^{K} p(k|\boldsymbol{x}) \int \log\left[p(\boldsymbol{x},z_j)\right] p(z_j|\boldsymbol{x},k)dz_j \\ &+ \text{const.} \end{aligned} \tag{3.8}$$

where $p(y,\boldsymbol{x},z_j) = \sum_{k=1}^{K} p(k)p(y|k)p(\boldsymbol{x},z_j|k)$ and $p(\boldsymbol{x},z_j) = \sum_{k=1}^{K} p(k)p(\boldsymbol{x},z_j|k)$. Thus, computing G_j for all I features amounts to computing $Q = I * (C+1)*K$ one-dimensional integrals over Gaussian measures $p(z_j|\boldsymbol{x},k) = \mathcal{N}(\mu_j(\boldsymbol{x},k),\sigma_j^2(\boldsymbol{x},k))$ with

$$\begin{aligned} \mu_j(\boldsymbol{x},k) &= \mu_{j,k} + \boldsymbol{\Sigma}_{z_j,\boldsymbol{x},k}\boldsymbol{\Sigma}_{\boldsymbol{x},\boldsymbol{x},k}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_{\boldsymbol{x},k}) \\ \sigma_j^2(\boldsymbol{x},k) &= \sigma_{j,k}^2 - \boldsymbol{\Sigma}_{z_j,\boldsymbol{x},k}\boldsymbol{\Sigma}_{\boldsymbol{x},\boldsymbol{x},k}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{x},z_j,k}. \end{aligned} \tag{3.9}$$

In these expressions $\mu_{j,k}, \sigma_{j,k}^2$ are the mean and variance of the jth feature in the kth component, while $\boldsymbol{\Sigma}_{a,b,k}$ is the part of the covariance matrix of the k component corresponding to variable sets a, b.

In order to be able to simulate strong and weak top-down attention, we may augment the model by smoothing the label-component table:

$$p(y|k) \to p(y|k,\beta) = \frac{p(y|k)^\beta}{\sum_c p(y|k)^\beta}, \tag{3.10}$$

and by letting the attention selection be stochastic based on the expected

gains, i.e., we select attention using the induced probability distribution,

$$P(j) = \frac{\exp(\gamma G_j)}{\sum_j \exp(\gamma G_{j'})} \tag{3.11}$$

Then, strong and weak task can be simulated changing the values β and γ.

3.3.3 Experiments with Human Subjects: The Role of Top-Down Attention in the Cocktail Party Problem

3.3.3.1 Experiment Design

We designed behavioral auditory experiments to investigate the role of top-down attention in a cocktail party problem taking Cherry's work in [Che53] as a reference. We analyze the human ability to attend to a single speech in a monaural mixture of two speech signals. We use two different narratives for each experiment. The texts are reported below [Mar12]:

- First Story: The giant panda still lives in the wild in only a few mountain ranges in the southwestern part of China because its survival has been threatened. both by hunters and by the destruction of the habitat it needs to survive. What has been noted and stressed in the last few decades is that the pandas survival is also threatened by the reproduction cycles of the bamboo where the pandas live. Here's what the problem is. It is the main source of food for the giant panda. However, when there's a massive reproduction of the bamboo, the one that has just reproduced dies, so there's a lag of quite a few years before the new, young seedlings grow enough to provide food. If the bamboo where the giant pandas living dies, then the giant panda needs to move to new areas to find food. The search for food has led the giant panda into areas, that are more settled and more full of danger.

- Second Story: Conifers are hardy trees that have been able to live long, so, as a result, both the oldest and the biggest trees in the world belong to the conifer family. The oldest known tree is located in east California. That tree is a four thousand years old bristlecone pine. The giant redwoods, in California, are the largest and oldest trees; they can be several hundred feet tall with a weight of two thousand tons. An interesting note about the giant redwoods is that, even though the trees are so big and tall, they have small cones. They are evergreens with short and spiky leaves. The needle-like shape of Conifer leaves evolved as a reaction to drought and aridity. When compared with a flat leaf, a needle presents a much smaller surface area. Most conifers are evergreens. They lose and replace their needles throughout the year, rather than shedding all their leaves in one season.

We reduced spatial and speaker cues, so that the solution is based on high level features related to content and task which we looked for in this work. Two speech signals were created with a speech synthesizer by AT&T Labs [BCS+99] (same speaker), so that the listener would not prefer one speaker over one. Considering the loudness, we equalized the long term loudness of the two speech signals (word by word loudnesses could still differ). In order to reduce the semantic masking effect, so that the listener would not attend to one story cause it is more interesting (at least as much as possible), we tried to choose neutral texts (not emotional etc.). We used texts in [Phi06] that is a preparation book for the TOEFL (Test of English as a Foreign Language) test . We made minor changes in the stories by adding and extracting some sentences and punctuation marks, changing the order of the words; in order to avoid the *listening in gaps* effect, described by Bregman [Bre94a] and Bronkhorst [Bro00], so that the listener would not take the advantage of pauses to switch attention.

We had two sets of experiments. In the first set, that we call *undirected attention experiments (UNDIR)*, the subjects are not assigned any task (they follow any story at anytime). In the second set, that we call *directed attention experiments (DIR)*, they are asked to follow only one of the stories (they can decide which story to follow though). In both sets of experiments, the listeners are presented a list of words once they finish listening, to report the words they heard. The list contains 48 words in

total, 24 of which occurred in the stories (12 for each story). We recruited 12 participants who were students. The participants were made to listen to the monaural mixture of two speech signals and report the words they heard afterwards; they repeated this process 3 times.

3.3.3.2 Temporal and Spectral Overlap

We define overlap using the *ideal binary mask (IBM)*, which has been attributed as the goal of computational auditory scene analysis (CASA) that studies perceptual audition [Wan05a], as already been studied in Chapter 2 in detail. We compute IBMs of the words that are pronounced in any of the stories and are also in the word list of the experiments presented to the participants in the end. We also compute the IBMs of the co-occuring parts (same time frame) in the second story.

Once we have IBMs for both speech signals (a word pronounced in one story and the corresponding part in the second story), we simply compute the percentage of the overlapping ones on both masks over the whole time-frequency bins and call it the *spectral overlap*. Then, we check each time frame and if there is at least one bin that has a value of one, we assign one for that time frame and we call the resultant map as the *compressed ideal binary mask (CIBM)*. Once we finish for every time frame for both IBMs, we compare for both signals and overlapping ones represent the *temporal overlap*. The illustration of how temporal and spectral overlaps are computed using IBMs can ben seen in Figure 3.1.

3.4 Experimental Evaluation

First, we evaluated our computational attention model with synthetic data we created and datasets from UCI machine learning repository [FA10]. We created two different scenarios; in the first one, the GDMM is trained with complete data while tested with missing data. In the second scenario instead, the training data is missing as well making the scenario more realistic (it is hard to get complete training data in real life situations). We used CEM as the missing data technique to deal

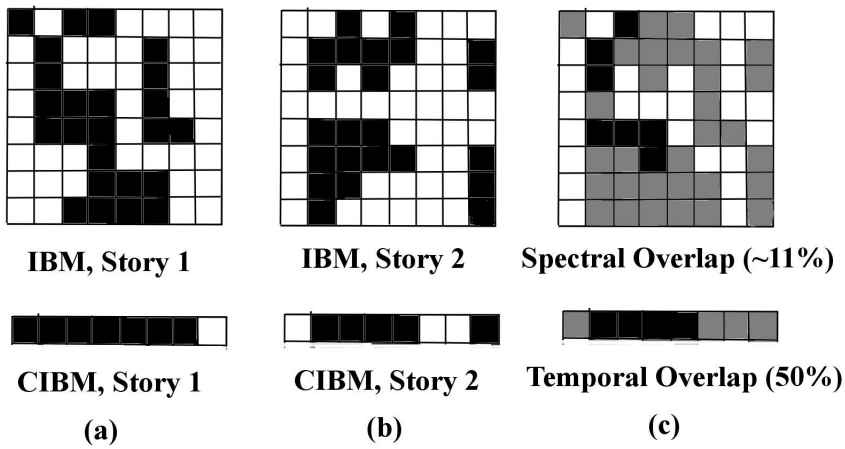


Figure 3.1: The illustrations of temporal and spectral overlap definitions, the bins represent time-frequency regions of an IBM. Only black regions represent overlapped parts on (c).

with missing data which is explained in section 3.3.2.2. We had similar results for both scenarios, and we will include only the results of the second scenario here to discuss the results. The results of the first scenario can be found in Appendix C.

Data is missing completely at random (MCAR) with different missing data percentages. The randomization is done such that not all values can be missing in one observation (one feature per observation is kept present randomly). As stated before, we look for the next feature to attend among missing ones using our attention model. We test the performance either choosing the next feature to be measured with the highest saliency or randomly and compare with classifiers, one trained with full data and the other the with original missing data we have. Since, we would like to reduce the error rate of an ensuing decision classification problem making use of our attention model, we evaluate its robustness again by the misclassification rates on test data.

Synthetic data set is generated by a Gaussian mixture model to test the algorithm. The number of mixtures K, is 3, and the number of

dimensions D is 10. The difficulty of the problem is determined using the SNR calculation below.

Let ds_{kl} be the distance between μ_k and μ_l, eig_k be a vector consisting of eigenvalues of Σ_k and mean() be the arithmetic mean operator, then

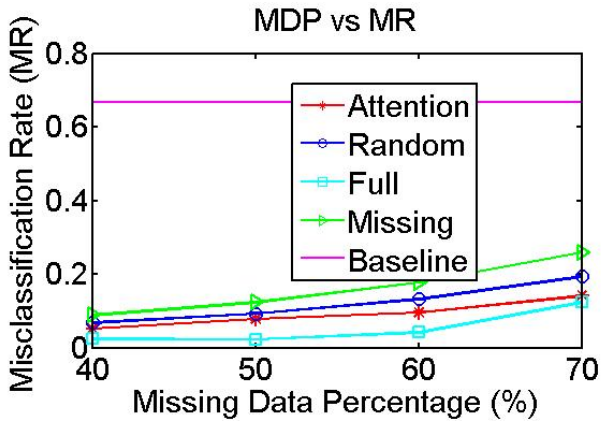
$$\mathrm{SNR_{dB}} = 10 \log \left(\frac{\sum_{1 \le \mathrm{k} \le \mathrm{K}, \mathrm{k} \le \mathrm{l} \le \mathrm{K}} \mathrm{ds}_{\mathrm{kl}}^2}{\sum_{1 \le \mathrm{k} \le \mathrm{K}} \mathrm{mean}(\mathrm{eig_k})} \right) \tag{3.12}$$

We use SNR of 10 dB, for a 10 dimensional data. 1500 and 300 observations are generated for training and test sets respectively.

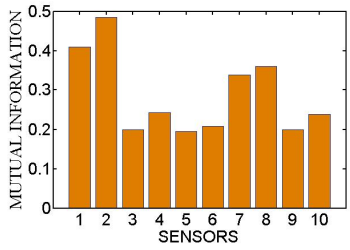
We designed the data set such that all features are equally important, which can also be seen in Figure 3.2 (b). We measure the mutual information between each input feature and the class label. We use a permutation test (Nresamples = 200) to test the mutual information against a null hypothesis of no mutual information. Mutual information is recorded if the null is rejected with $p > 0.01$. Figure 3.2 (c) shows the frequency of which features are selected with the attention model. We observe that the frequencies are not simply given by the mutual information showing the need for an attention model. Figure 3.2 (a) shows the error rates for different missing data percentages. As expected, we reduce the error rate adding one feature randomly or with attention model; yet, the top-down attention model outperforms the random saliency model.

Another data set used is the Yeast dataset, used for illustration of the top down saliency estimate, concerns determination of protein cellular localization sites [HN96].

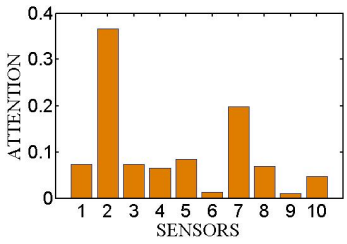
1. McGeoch's method for signal sequence recognition
2. Von Heijne's method for signal sequence recognition
3. score of the ALOM membrane spanning region prediction program
4. score of discriminant analysis of the amino acid content of the N-terminal region (20 residues long) of mitochondrial and non-mitochondrial proteins



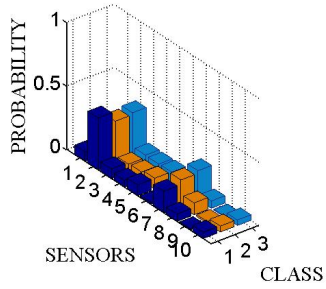
(a)



(b)



(c)



(d)

Figure 3.2: Synthetic data. Ten input features are considered. Train and test data are missing completely at random (MCAR). (a) For different missing data percentages, the misclassification error rates for the test set where data is MCAR (Missing), all missing features are added (Full), one random feature is added among missing (Random) and one feature is added among missing using the attention model (Attention). (b) The log_2 mutual information between features and class label. (c) Frequency of selection of additional features with attention model. (d) Frequency of selection of features within the three classes.

5. presence of 'HDEL' substring (thought to act as a signal for retention in the endoplasmic reticulum lumen)
6. peroxisomal targeting signal in the C-terminus
7. score of discriminant analysis of the amino acid content of vacuolar and extracellular proteins
8. score of discriminant analysis of nuclear localization signals of nuclear and non-nuclear proteins

We reduce the classification problem to a binary decision by selecting a subset associated with two most frequent sequence types CYT (cytosolic/cytoskeletal 463 examples) and NUC (nuclear, 429 examples) in SWISS-PROT database. We have a a training set with 630 samples, while a test set with 262 samples. We train the Gaussian Discrete model with K = 11.

As seen in Figure 3.3 (b), features (1,8) are the most informative ones. Feature 8 is also the most selected feature as seen in Figure 3.3 (c), yet, features that are not informative are selected often as well showing again the need for an attention model. Figure 3.3 (a) shows the error rates for a missing data percentage of 50% and as for the sythetic data set, we observe that the attention model outperforms the random saliency model.

We simulated also the effect of strong task (DIR) and weak task (UNDIR) in top-down attention, as explained in section 3.3.2.3, by smoothing the label-component table as in Equation 3.10 and by letting the attention selection be stochastic based on the expected gains as in Equation 3.11. We confound the input test signal; we replace a fraction $f(overlap)$ of the input signal with a randomly chosen input feature vector from the opposite class (confounder).

$$confounded = (1\text{-}f)^* \ input\ signal + f\ ^*\ confounder$$

The C = 2 simulated decision problem is based on a four component Gaussian mixture. The 2-dimensional gist is provided and an additional

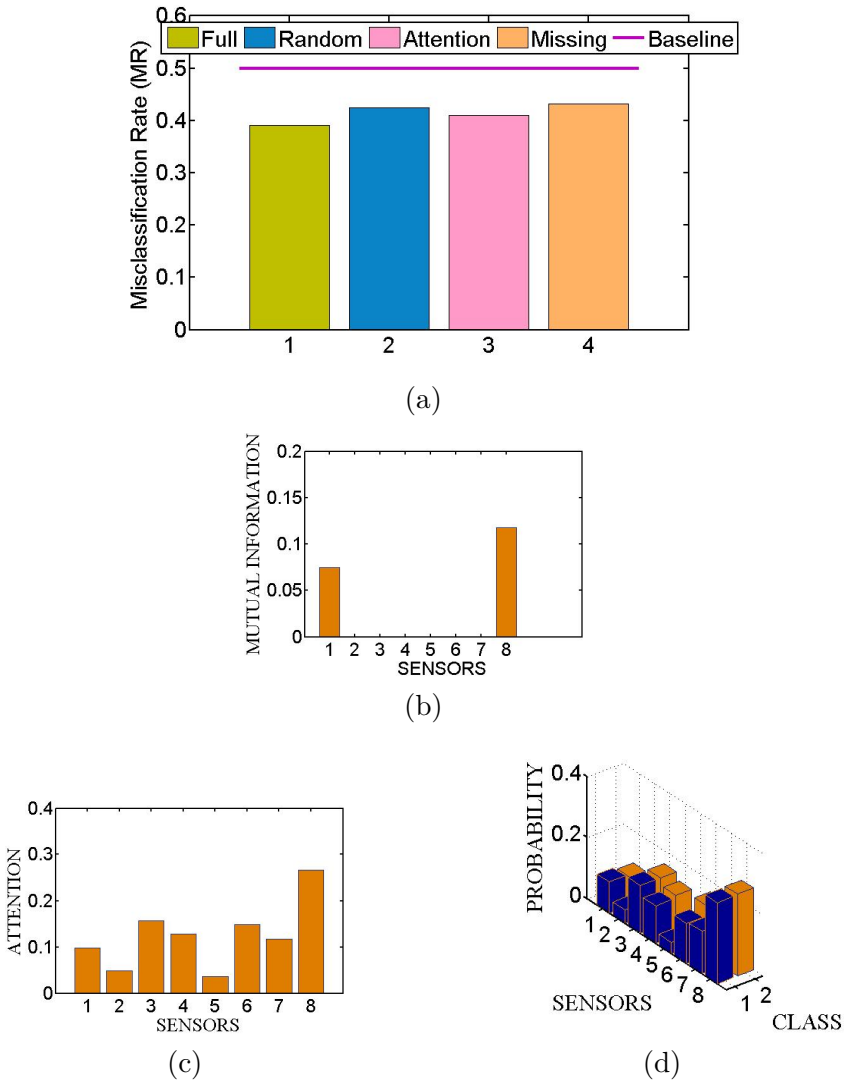
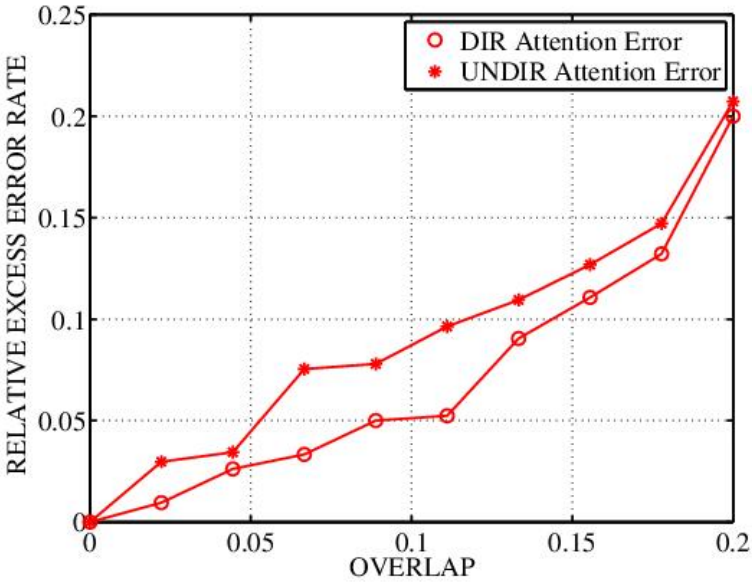


Figure 3.3: Yeast data. Eight input features are considered. Train and test data are missing completely at random (MCAR). (a) For 50% of missing data percentage , the misclassification error rates for the test set where data is MCAR (Missing), all missing features are added (Full), one random feature is added among missing (Random) and one feature is added among missing using the attention model (Attention). (b) The log_2 mutual information between features and class label. (c) Frequency of selection of additional features with attention model. (d) Frequency of selection of features within the two classes.

feature among the six remaining features is selected using the attention mechanism. The error rates are represented as relative excess errors: $[E(f) - E(0)]/(B - E(0))$, where $B = 0.5$ is the baseline error rate. The error rate of the strong task attention model is $EDIR(0) = 0.08$ while the error rate of the weak task attention model is $EUNDIR(0) = 0.23$. We observe that strong task attention model (DIR) is less sensitive to the confounding mixture than the weak task attention model (UNDIR) as seen in Figure 3.4, hence it will make more informed decisions in the cocktail party.



(a)

Figure 3.4: The resulting error rates for the models using the attention mechanism to select an additional feature among the six remaining after having provided the 2-dimensional gist for a range of mixing fractions f [0.0 0.2].

Finally, we give the results of the behavioral experiments here. Figure 3.5 and Figure 3.6 give the temporal and spectral overlaps for UNDIR and DIR cases respectively. The IBM parameters used are the ones given in previous works for high speech intelligibility results in studies [KBP+09a] by Kjems et.al. and [KLPB10b] by us. The correlation between over-

laps and the number of times the words heard are -0.35 and -0.31 for UNDIR case. As expected, fewer times a word is heard if the overlap is high. However, in DIR case, we are not able to observe the same, instead we observed positive correlation. We also optimized the IBM parameters for these cases (the optimization process can be seen in detail in Appendix E). The correlation we found for DIR case with optimized parameters were negative (-0.2 and -0.03), yet, still much lower than for the UNDIR case (-0.59 and -0.43). In both spectral and temporal overlaps, for UNDIR experiments,under the 5% significance level, the null hypothesis that the data is uncorrelated is rejected, while accepted for DIR experiments. We conclude that the relation between overlap and speech intelligibility is modulated by the presence of a task, hence top-down controlled attention.

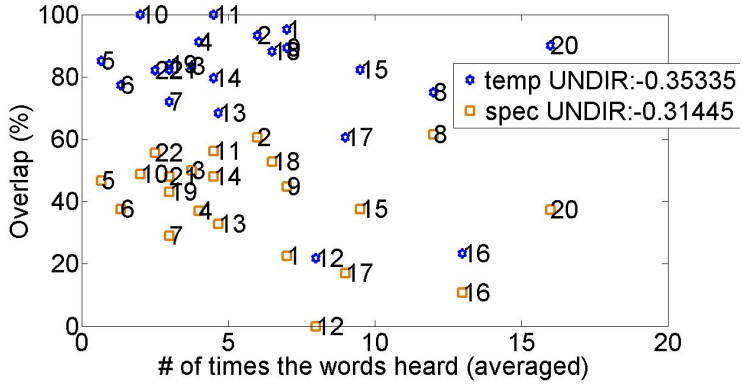


Figure 3.5: Temporal and spectral overlap versus the averaged number of times the words heard for UNDIR case, and the correlation between them shown on the legend.

3.5 Summary

We proposed a top-down attention model based on a sequential decision making. The saliency of features was given as the difference in confusion (entropy). By using a Gaussian mixture as the learning model we obtained a relatively simple expression for the feature dependent information gain. Both train and test data were assumed to be incomplete. We

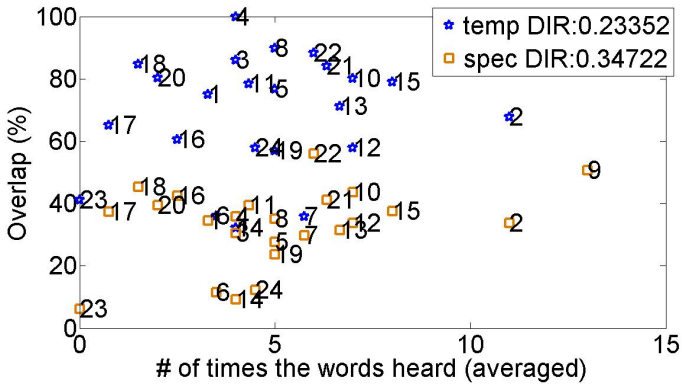


Figure 3.6: Temporal and spectral overlap versus the averaged number of times the words heard for DIR case, and the correlation between them shown on the legend.

applied the method to two classification problems (synthetic and Yeast datasets) and we showed that the top-down attention model we proposed outperformed simple random saliency model.

We also simulated the cocktail party effect based on our recent top-down attention model. We showed that strong task attention model showed less insensitivity to the amount of the confounding overlap, than the weak task attention model. We conclude that the top-down attention can undermine the effect of overlapping noise.

In the 'hard cocktail party' behavioral experiment, with reduced spatial and speaker cues, we observed significant negative correlation between the number of times words are heard and temporal and spectral overlaps for UNDIR case (no task, follow the sound signal). Meanwhile, for DIR case (with task, follow one of the stories), no significant correlation was observed. We conclude that top-down knowledge can enhance speech intelligibility by compensating for noise and the presence of confounders.

Chapter 4

Automatic Emotion Recognition from Speech

This chapter describes a dimensional speech affect recognition model that combines acoustic and semantic features. Sections 4.1 and 4.2 state background information, motivation and related work in the literature. Section 4.3 describes the method used to design the model and explains and analyzes the dataset obtained to evaluate the model. Section 4.4 reveals the experimental results and discusses these results. Section 4.5 summarizes the chapter.

4.1 Background

Darwin addressed the emotion theory from an evolutionary point of view [Dar65]. He observed that emotions prime behaviors that enables others to conceive one's emotional state, and suggested that emotions are an evolved communication mean. Vocal characteristics, facial displays, and whole-body behaviors of a person usually convey one's emotional state

[MR09]. Humans acquire the intuition of reading one's mind, or emotional state using cues from the former characteristics from early ages, thus, it is a quite easy task intuitively. However, the theory behind how humans measure emotions is still in search in affective science.

There are two main emotional modelings approaches which are categorical and dimensional approaches. In categorical approach, the emotions are classified discretely, such as happiness, sadness and surprise. In this approach, the emotion theory claims that the emotions are hard-wired in the brain and recognized universally [GP10]. Researchers usually refer to 'basic emotions', that are anger, fear, disgust, happiness, sadness, and surprise and 'basic emotions' have been shown to be recognized universally by facial behaviors by Ekman in [EF71] following Darwin's work in [Dar65]. The two meanings of the word 'basic' in this context are defined by Ekman in [Ekm92] as that emotions differ from each other not only in expression but also in other aspects such as appraisal, probable behavioral response, physiology, etc. and that they evolved dealing with fundamental life-tasks.

Some researchers argue that emotional states are not independent from one another but related in a systematic manner and they model the emotion with a dimensional approach. The most commonly assumed dimensions are valence, arousal, dominance-submission and approach-avoidance [LBC$^+$97, BR99]. Dominance-submission is a continuum ranging from feeling influential and in control to the opposite extreme feeling of lack of control or influence. The approach and avoidance motivations describe the tendencies to approach and avoid stimuli respectively. The arousal dimension describes how excited one is, or how much energy is required to express the feeling. Feelings with high arousal induce some physical changes in the body such as increased heart rate, higher blood pressure and greater sub-glottal pressure resulting in change in speech as well such as making it louder, faster and have higher average pitch etc. The valence dimension describes to which extent the feeling is positive or negative, from unpleasant to pleasant. It has not been shown yet, how or if the acoustics features correlate with the valence dimension [EAKK11], and it has been claimed to be harder to estimate automatically compared to the arousal dimension [HGSW04]. Valence and arousal dimensions have been often preferred by researchers since they have been

shown to cover the majority of affect variability [GP10, Fer04].

4.2 Related Work

Emotion recognition from speech has attracted many researchers in both dimensional and categorical approach [NFDS03, VK05, PY03, SRM+05, LN05, YTCZ11, EWG+10, ZZL11, Sch11, FT05, LCCS07]. Most of the emphasis has been on the categorical approach until recently, classifying emotions discretely into mostly 'basic' emotions (anger, fear, disgust, happiness, sadness, and surprise) [NFDS03, VK05, PY03, SRM+05, YTCZ11, ZZL11] or negative-positive feelings [LN05]. Recently, the focus on speech emotion recognition has been shifting towards the dimensional approach in which usually valence and arousal dimensions are recognized [EWG+10, Sch11, FT05, LCCS07].

Speech emotion recognition is a challenging research area and one of the main challenges is the extraction of a proper feature set [EAKK11]. Unfortunately, there is not a standard feature set proved to characterize the emotional content effectively independent of the speaker, environment or the lexical content. Meanwhile, the typical acoustic features applied in this area are the pitch, the formants, the short-term energy, the mel-frequency cepstral coefficients (MFCC), and the Teager energy operator based features [Ver06, NFDS03]. Hundreds of acoustic features can be gathered using the speech signal which could lead redundant information and computational burden, which also constitutes another important challenge in this area. In many cases, feature reduction and selection methods are applied to deal with huge number of features [SRM+05, LN05, BSS+11, PY03, VA05]. Most of the used methods are information gain [PY03], Sequential Forward Floating Search (SFFS) [SRM+05] and correlation-based feature selection (CFS) [VA05]. Once the features are extracted and selected properly, the choice of an appropriate learning model is important as well. Many models have been used for this task, such as hidden Markov model (HMM) [NFDS03, PSB+10], support vector machines (SVM) [SRM+05, YTCZ11, HHDH09], K-nearest-neighbor (KNN) [YTCZ11], Gaussian mixture models (GMM) [KHKK09] and neural networks (NN) [EWG+10, ZZL11], but there is yet no agree-

ment on the most suitable one.

There have been effort also on combining acoustic and semantic information of speech and has been shown to improve the performance [SRL04, CW04, Sch11]. Bag-of-words (BOW) features, where each term within a vocabulary is represented by a feature modeled by the term's occurrence frequency within the phrase, and part-of-speech (POS) features, where each phrase is represented by grammatical tags (verbs, nouns, etc.), have been shown to be useful linguistic features for speech emotion recognition task [BSS+11, Sch11, BSS+06]. The vocabulary is often limited to predefined emotional keywords including words like 'happy', 'sad' and 'depressed' [CW04]. Semantic language analysis, which is the study of 'meaning' and is a subfield of linguistics, has been investigated as well, which could be used to enable to use not only predefined emotional keywords but also general terms that could carry the emotional content as well. Latent semantics analysis (LSA) is an indexing method that is based on the principle that words occurring in similar contexts are also similar in meanings, and it has been used in emotion recognition from speech or music [YNP11, PH10a].

Words play an important role in semantic analysis of speech as in using BOW and LSA methods and some studies have been carried out on how to evaluate each word in the emotional dimensions. The affective norms for English words (ANEW) in [BL99] introduced by Bradley and Lang in 1999, includes a set of normative emotional ratings for 1034 English words, in arousal, valence and dominance dimensions. It provides the researchers with the standardized materials in emotion studies. The self-assessment manikin (SAM), an affective rating systems designed by Lang in 1990 [BL94], is used to assess the three affect dimensions which are valence, arousal and dominance. ANEW has been widely used for emotion analysis purposes by researchers and has also been adapted for different languages [SMJ07, LCRD07, RFPC07]. Figure 4.1 shows the 'basic' emotions mapped onto 2 dimensional (arousal and valence) space using arithmetic mean values for the ANEW words 'anger', 'fear', 'surprised', 'disgusted', 'sad' and 'happy'.

Last, the choice of a proper database is another challenge. There are many aspects to be considered while trying to choose a database, such as

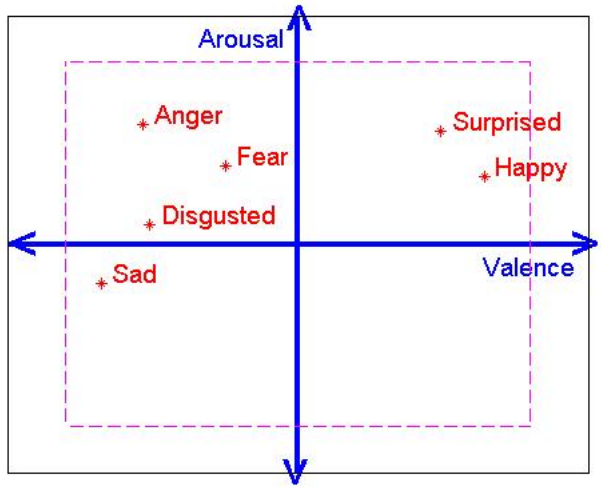


Figure 4.1: A-V space with 'basic' emotions mapped onto it using the corresponding ANEW words' mean A-V ratings

the language, the scope (emotion analysis or recognition), subjects and observers (adults or children), naturalness (acted or natural), the balance in phrases (the number of phrases per emotion, phrase length, etc.), the emotion model (categorical or dimensional), the assessment type (which emotions for the categorical, continuous or quantized for dimensional, etc.) and the duration of phrases or dialogs, etc.[1]. Although, there are a number of available databases developed well, it is usually a challenge to find a suitable one for a specific purpose. Therefore, researchers may prefer to design their own databases [CW04, LW05].

[1] [EAKK11] can be checked for a review of some commonly used emotional speech databases

4.3 Method

4.3.1 Database Design and Analysis

4.3.1.1 Purpose and design criteria

The purpose of our work has been to combine semantic and acoustic information of the speech signal to recognize emotions in arousal and valence dimensions. For this purpose, we used ANEW words, that are rated on A-V space by human observers, to extract semantic information. This step of our model brought some constraints on the database we needed as seen below.

- Language: English
- Emotion Model: 2 dimensional model (valence and arousal)
- Response format: Similar to ANEW format as much as possible
- Annotators: Adults
- Content : Speech including words with emotional content (not only neutral words like date, time etc.)

There are a few well designed databases for speech emotion recognition subject [EAKK11, DCCCR03], however most of them were designed for the categorical approach [Uni02, BPR+05, EH96, HKM+02, SRM+05]. Although, the dimensional approach is in its infant stage, there have been attempts on designing databases using this approach as well [GKN08, DCCS+07, BFBG08, MVCP10]. However, in [GKN08], the language is German, in [DCCS+07], the scope is emotion analysis rather than recognition and in [BFBG08], the content consists of nonverbal affect bursts.The Semaine database in [MVCP10], has been the closest to satisfying our needs, yet, they use a different response format (no use of SAM figures, continuous and a rating scale between [-1,1]). Additionally, one of our objectives in this study has been to investigate the effects of semantic and acoustic information on arousal and valence dimensions

separately. In the database we designed, the clips are rated either just reading the text part or just listening to the audio part giving us the opportunity that Semaine database lacks, to analyze the effect of semantic and acoustic features on affective dimensions.

4.3.1.2 Method

The database consists of 59 sections from 11 movies in total. The duration of each section differs between 5 and 25 seconds. The difficulty about selecting the sections was the background noise, that is usually music that is used in the movies usually to induce the feeling that is expressed in that particular section more. We took the sections without this kind of noise, including monologues or dialogs. We will refer to these sections as 'clips'. Selecting the clips, we tried to have ratings in all four quadrants in A-V space, which are high arousal low valence (HL), high arousal high valence (HH), low arousal low valence (LL) and low arousal high valence (LH) based on the author's judgments initially.

The clips include audio and text in the form of subtitles. The clips were resampled at 16 kHz. The long-term loudnesses of the clips have been normalized using Replay Gain[2], which is an open standard loudness calculation algorithm [Rob01] in which the main idea is to calculate the gain needed on an audio file to match the perceived loudness level of a reference audio file.

The response format used creating this database is the same one used in ANEW work. This choice was based on the fact that we aimed to coincide with ANEW work that we use in our learning model. The clips were rated in valence and arousal dimensions in a discrete way using SAM figures [BL94]. Figure 4.2 shows the SAM figures we used in our work.

A Java applet was designed for the experiments to obtain the ratings and was available online for the participants. There are 3 experiments and a questionnaire to be filled in by the participants, which take around

[2]Please note that the instantaneous loudness in the clips can still be used as an acoustics feature, the normalization is intended just for long-term loudness

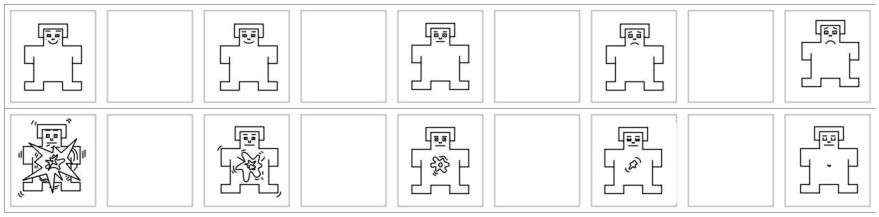


Figure 4.2: The SAM figures [BL94] used to rate the clips in the database. The upper figures are to rate the valence from 'pleasent' to 'unpleasent', while the bottom figures are for the arousal from 'excited' to 'calm'.

one hour in total to finish. The applet design can be seen in Figure 4.3. The users start with filling in parts to gather some personal information such as age, occupation and gender for statistical purposes. Once all the necessary information is filled in, experiment 1 is enabled. In the first experiment, just the text (JT) parts of all 59 clips are displayed one by one, and the users rate each on the A-V space using the SAM figures. Once, they finish rating all the clips, the second experiment is enabled in which, the users again rate all the clips, but just with the audio (JA) part, without the text part. Later, the experiment 3 is enabled in which they rate all the clips with both text and audio parts. Before each experiment, the order of the clips are randomized. Finally, the users are asked to fill in a questionnaire about the experiments again for statistical and feedback purposes.

The user is provided with the document including the instructions, which is provided in [Kar12], before starting the experiments. The rating format used can be seen in Figure 4.4. There are the SAM figures that can be clicked on to rate the feelings in a clip, five buttons ('Start','Previous','Next','Play Again','Finish') and the 'black' part in which the transcription appears synchronized with the audio in the clips. The user is enabled to listen to a clip more than once, which could be beneficial in cases that he/she could not hear well, or was disturbed, etc. They are also enabled to go back to the previous clip so that they could change their ratings in case they made a mistake rating or were not sure about their ratings etc.. In each experiment, first two clips are provided just for training purposes.

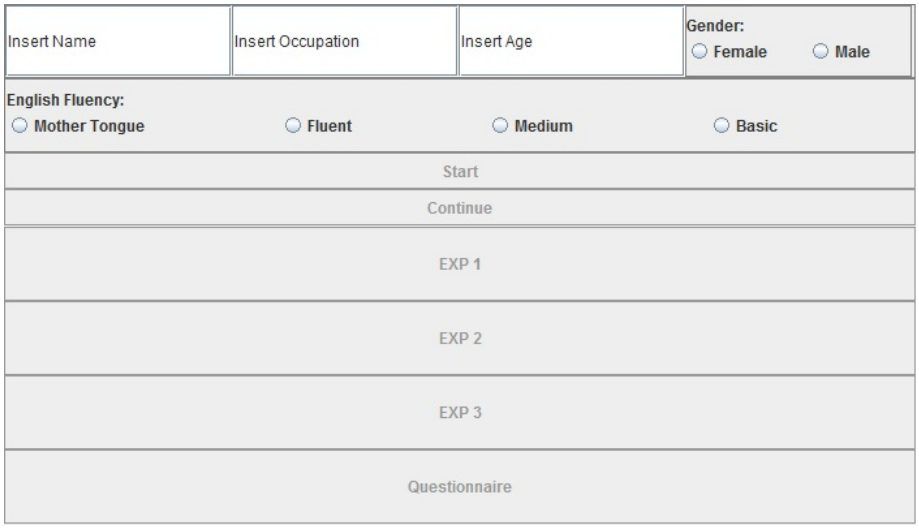


Figure 4.3: The applet used to design the database. There are personal information parts to be filled in by users before starting, 3 experiments (EXP1 (JT), EXP2 (JA), EXP3 (BTA))and a questionnaire about the experiments.

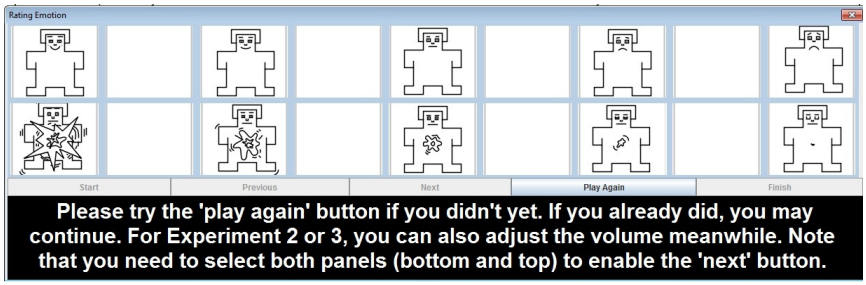


Figure 4.4: The rating format used for the database. The text part of the clips appear on the 'black' part synchronized with the audio part. The text seen in this example does not belong to any of the clips, but used for training purposes.

13 people, (7 female, 6 male), between ages 19 and 28, speaking English fluently, were recruited. The participants were able to carry out the experiments from a place they prefer as long as they had Internet with a reliable connection. They were asked to do it in a not disturbing place using a headphone and not to give long breaks between the clips. The time track of each user's rating has been collected through the applet and these recordings were checked for consistency, for existence of any long breaks, etc.[3]. In the questionnaire, all users reported that the instructions were clear and they were confident in rating. The Java code for the applet[4], the start and end times of the clips in each movie (without the audio clips themselves due to copyright restrictions) and the ratings obtained from all participants are provided in [Kar12]. We call this database as Emotional Movie Database (EMOV).

Table 4.1 shows the properties for the EMOV database obtained.

Emotion Model	2 dimensions, Arousal and Valence
Content	59 short clips from 11 movies in total monologues/dialogues
Duration	5 to 20 seconds
Language	English
Audio	16 kHz, 16 bits
Text	421 words in total (after stop words are extracted)
Annotators	13 people (7 female / 6 male,between ages 19 and 28)
Response Format	Rated in A-V space, quantized between 1 and 9

Table 4.1: EMOV Database properties

[3]In case of failure to finish all three experiments due to technical difficulties or other reasons, the results have not been included

[4]The code could be used to replicate the experiments, or add more data, or make similar experiments with different clips, etc.

4.3.1.3 Analysis

We applied Peirce's outliers detection algorithm [Ros03] to the data to reject outliers and 2.9% and 3.7% of data per person was rejected from valence and arousal dimensions respectively. Figure 4.5 shows the arithmetic mean of ratings on arousal and valence dimensions, with outliers rejected for the three experiments using JT, JA and BTA. We have ratings in four quadrants that are HH,HL,LH,LL, although not balenced in all (more data on HL).

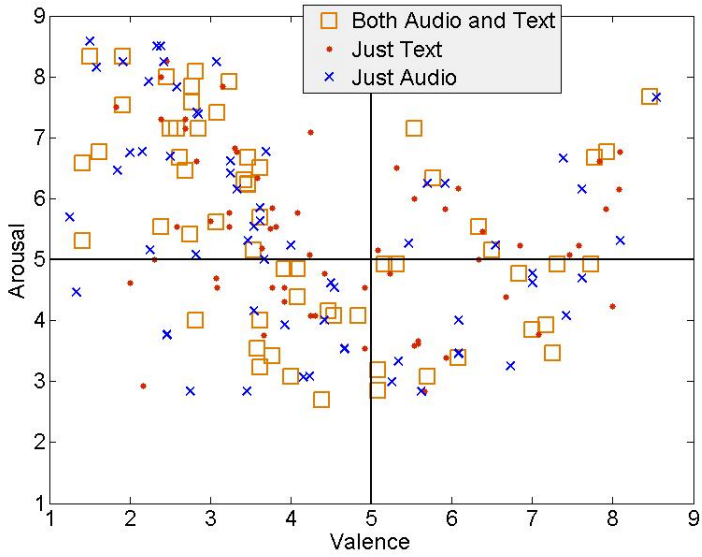


Figure 4.5: The arithmetic mean ratings over people of all the clips for the three experiments using JT, JA and BTA.

Figure 4.6 and 4.7 show the distributions for all three experiments, for valence and arousal dimensions respectively. It is observed that the distributions of the results of the experiments differ more between using JT and BTA compared to JA and BTA. It is also observed in Table 4.2 which shows arithmetic mean of the differences in the ratings of the experiments using JT and BTA, and JA and BTA. Using JA and BTA do not differ that much since the users speak English fluently, and by listening to the clips they already could get the semantic information as

well. However, we observe a considerable difference between using JT and BTA. We also observe that this difference (between JT and BTA) is bigger for the arousal dimension than for the valence dimension, which we infer as that rating the valence dimension is easier than rating the arousal dimension using just semantic information from a speech signal.

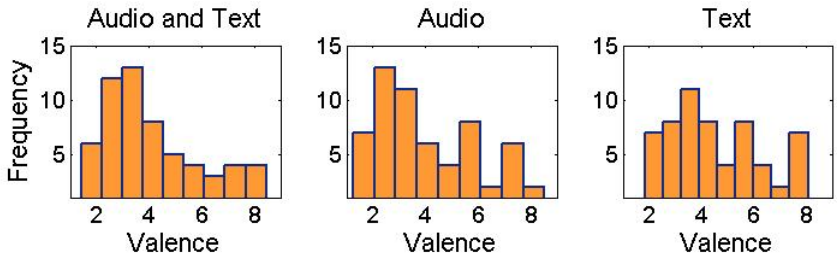


Figure 4.6: The distribution of ratings for 59 clips averaged over people for three experiments for valence dimension.

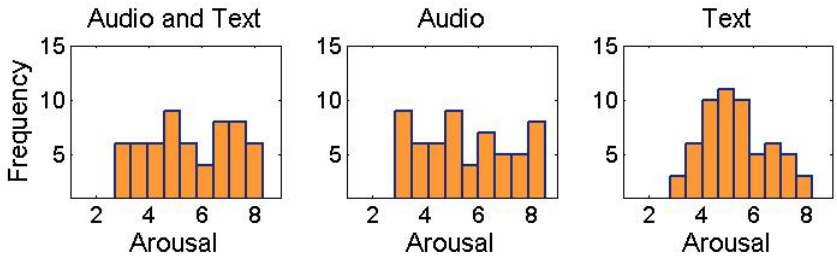


Figure 4.7: The distribution of ratings for 59 clips averaged over people for three experiments for arousal dimension.

4.3.2 Modeling Framework

4.3.2.1 An Overview:

The overview of the modeling framework is given in Figure 4.8. The recognition of the emotions is carried out separately for acoustic and semantic part and then later the results are fused in, in order to obtain the final results. In semantic part, LSA is applied on the transcriptions

	Valence	Arousal
JT-BTA	0.62	0.86
JA-BTA	0.26	0.35

Table 4.2: The arithmetic mean of the differences in the ratings of the experiments using JT and BTA, and JA and BTA

of the audio clips (text part) to extract the similarity of each word in a clip to ANEW words. Then, combining these similarity scores and A-V values of ANEW words, each word's A-V values are estimated, and later these estimations for all words in a clip are combined and the resultant values are attributed as the clip's A-V values. In acoustic part, feature extraction, selection and regression (using SVR) steps are followed in order. Finally, the results of acoustic and semantic parts are combined using the equation below:

$$AV_{comb}(i) = ws(i) * AV_{sem}(i) + wa(i) * AV_{aco}(i) \tag{4.1}$$

where, $AV_{comb}(i)$, $AV_{sem}(i)$, $AV_{ac}(i)$ represent the A-V values of the i^th clip for the combination, semantic and acoustic parts, and $ws(i)$ and $wa(i) = (1 - ws(i))$ are the weights of semantic and acoustic parts in the combination. The details of each part will be explained in the following paragraphs.

4.3.2.2 Acoustic:

Openear toolkit [EWS09], an open-source affect and emotion recognition engine, was used to work in all feature extraction, selection and regression parts. Low level audio features, such as signal energy, pitch, Mel frequency cepstral coefficients (MFCC), perceptual linear predictive coefficients (PLPC) and formants, are extracted and various statistical functionals and transformations, such as extremes, means (arithmetic, geometric, quadratic), moments (variance, skewness etc.), peaks (number of peaks, mean peak distance etc.), are applied to those features.

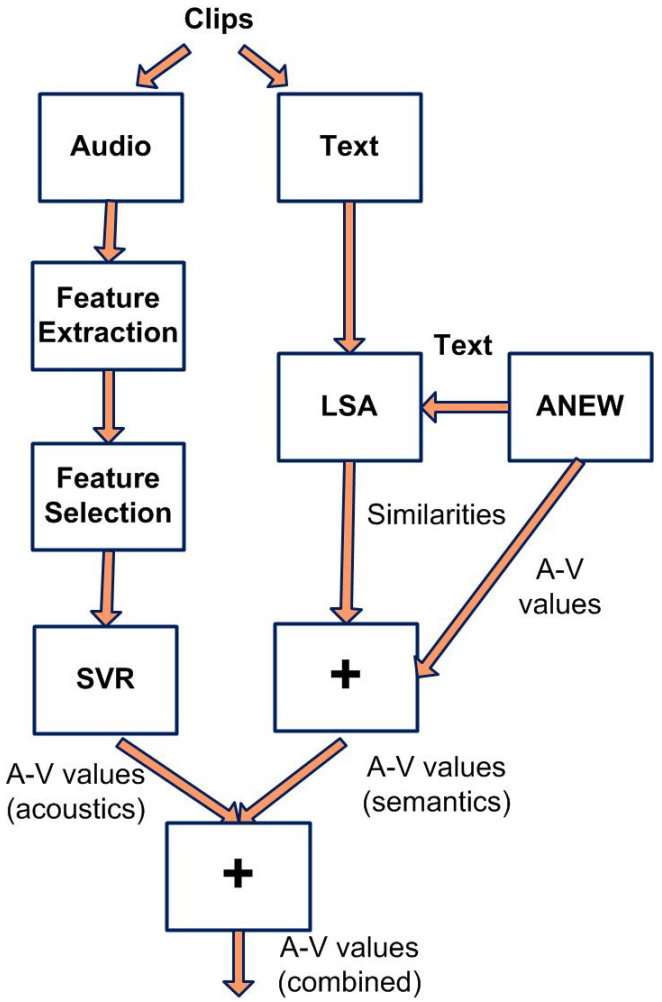


Figure 4.8: An overview of the framework. The (+) sign represents that the inputs are combined, the details of the combination methods are in the relative sections.

988 features were gathered in total. Correlation feature selection (CFS), which is based on the hypothesis that good features are the ones which do not correlate with each other but correlate with the classification, was used to select the features. Finally, we use support vector regression (SVR) [SS04] method to have the estimated A-V values.

4.3.2.3 Semantic:

We use the A-V ratings of 1034 ANEW words on valence and arousal dimensions to map the words in the clips on A-V space. We apply latent semantic analysis (LSA) to calculate the similarity scores of each word in a clip to those ANEW words. An LSA software package [DTU10], that is based on term frequency inverse document frequency (TF-IDF) weighting and that outputs cosine distance as the similarity measure, has been used. We simply treat 1034 similarity scores we had as the weight of each ANEW word's contribution in the estimation of A-V value of a word. In order to get rid of redundant contributions, especially from those with so low similarity scores, one of the options would be to use a threshold value for the similarity score and take those above it. However, with this approach, there is a trade-off. If we choose the threshold to be high, if a word is not similar to any ANEW word more than this threshold value, we would not be able to estimate the A-V value of it. On the other hand, in case of a low threshold, we could have so much redundant information for the cases in which a word is highly similar to some of the ANEW words (the extreme example would be to have a similarity score of 1 meaning that it is the same word as one of ANEW words, we could take into account other ANEW words as well getting further from the real A-V value of that word). Therefore, instead of using the similarity score directly, we define a weight for each word using Equation 4.2, which solves this problem by adjusting the threshold using the information of the maximum similarity score of a word to any ANEW word,

$$we_{ANEW}(ji) = sim_{ANEW}(ji)^{\left(\frac{th}{1+gamma-sim_{max}(j)}\right)} \tag{4.2}$$

where $we_{ANEW}(ji)$, th and $sim_{max}(j)$ are the weight for each ANEW

word for the j^{th} clip word, a threshold value to be optimized between 0 and 1 and the maximum similarity of j^{th} clip word to ANEW words respectively. *gamma* is a very small number (close to zero) used to avoid dividing by zero in the case of sim_{max} of 1. Figure 4.9 illustrates the weight-similarity relation visually. The purple line on the figure, which gives weight of zero to all similarities below zero, shows the case when the maximum cosine similarity of a word to ANEW words is 1. In this case, only the A-V values of the corresponding ANEW word is taken as desired.

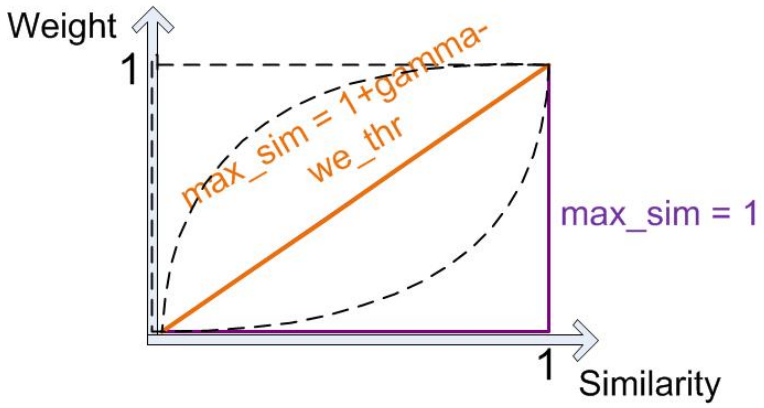


Figure 4.9: Similarity to weight conversion using Equation 4.2. The curve shape changes with th and sim_{max}. The purple and orange colored curves show two different specific cases.

Equation 4.3 is used to estimate A-V value of each word.

$$AV_{word}(j) = \frac{\sum_{i=1}^{N} AV_{ANEW}(ji) * we_{ANEW}(ji)}{\sum_{i=1}^{N} we_{ANEW}(ji)}, \tag{4.3}$$

where $AV_{ANEW}(ji)$ and N are the A-V values of the ANEW words for the j^{th} clip word to be analyzed and the number of ANEW words which is 1034, respectively. Finally, in order to estimate the A-V values of a clip, we use the estimated A-V values of each word and calculate the weighted average, taking the maximum similarity score found for each word as its weight.

4.4 Experimental Evaluation

The database we designed has been divided into a train set with 29 clips and a test set with 30 clips. The optimization of the parameters for the text part and the combination part has been done using the train set, and the final results are evaluated using the test set. We used a subset of ANEW words, the list of which can be reached in [Kar12]. More details on how experiments were carried out and the optimization process can be found in Appendix F.

The error between the estimated and human-rated A-V values is measured using mean absolute error (MAE) which gives the average of the absolute differences between them and root mean squared error (RMSE) which is a useful error measure for the applications where large errors might be specifically undesirable[5].

Table 4.3 and 4.4 show the resultant errors of semantic and acoustic parts respectively. It is observed that valence is better estimated using semantic information, while arousal is better estimated using acoustic information. We used Equation 4.1 with optimized *ws* and *wa* values, to have the results for the combination of acoustic and semantic parts and Table 4.5 shows the resultant errors of that combination. It is observed that by combining the two parts, we were able to increase the recognition performance slightly. In addition, and more importantly, we observed that for the combination of acoustic and semantic information to estimate A-V values, the semantic information should be weighted much higher to estimate valence, while the acoustic information should be weighted much higher to estimate arousal. We conclude that *the valence dimension is more about what we say while the arousal dimension is more about how we say it.*

[5]The errors are squared before they are averaged, so higher weight is given for large differences

	MAE	RMSE
Valence	1.45	1.85
Arousal	1.39	1.64

Table 4.3: Mean Absolute error (MAE) and Root Mean Squared Error (RMSE) for semantic part in A-V dimensions, using the test set

	MAE	RMSE
Valence	1.98	2.50
Arousal	1.29	1.54

Table 4.4: Mean Absolute error (MAE) and Root Mean Squared Error (RMSE) for acoustic part in A-V dimensions, using the test set

		Weights (ws / wa)	Combined Result
Valence	MAE	0.80 / 0.20	1.40
	RMSE	0.85 / 0.15	1.77
Arousal	MAE	0 / 1	1.29
	RMSE	0.20 / 0.80	1.52

Table 4.5: Mean Absolute error (MAE) and Root Mean Squared Error (RMSE) for the combination of semantics and acoustics parts in A-V dimensions with weight values of each part, using the test set

4.5 Summary

We investigated the speech emotion recognition fusing semantic and acoustic information within a dimensional approach. We investigated not only the effect of combining semantic and acoustic information, but also the effect of each part for different affective dimensions, which are arousal and valence in our work.

For the semantic part, we made use of ANEW words that are already

rated on the A-V space by humans which brought some constraints for the database we needed. Since we had many constraints about the type of the database we needed and since there are not many developed speech emotion databases for the dimensional approach, we designed our own database. The database consisted of 59 short movie clips with 5 to 25 seconds of duration, rated by human subjects.

We recognized the emotions using semantic and acoustic information separately. We showed that the integration of semantic and acoustic features improved the performance of dimensional speech emotion recognition and the weight of each part differed for valence and arousal dimensions. We concluded that *valence is more about what we say while arousal is more about how we say it.*

Chapter 5

Conclusion

Hearing aids have improved drastically, especially after the application of advanced DSP algorithms. Straightforward engineering applications driven by bottom-up approach, that are based on sensory and perceptual processes, have mostly driven the hearing aid technology. However, as human auditory system works also using top-down approach, that is about knowledge and based on cognitive processes, the focus of hearing aid technology has been shifting towards including top-down processes as well. The assumption, "one representation fits all", is another limitation of current hearing aids to satisfy the user needs. The needs of an individual may depend on many variables such as the physiology of hearing loss, cognitive abilities, mood, age etc., which should obviously be taken into account in the design of a hearing aid. Current hearing aids can to some extent adapt to different situations adjusting the hearing aid settings, for instance by classifying the environment (car, cafeteria etc.), yet they lack the ability to learn. We think that future hearings aids that we call 'cognizant hearing aids' will be designed as intelligent and individualized gadgets integrating top-down and bottom-up approaches. This goal will be reached by utilizing research done within cognitive science, that is the interdisciplinary study of mind and intelligence that brings together

many disciplines including Artificial Intelligence, Cognitive Psychology, Neuroscience, Linguistics, Anthropology and Philosophy.

This thesis has focused on three problems within cognitive science that would impact future hearing aids, which are automatic speech recognition using binary masks obtained by applying auditory scene analysis, modeling top-down auditory attention and analyzing human auditory attention in a cocktail problem and automatic emotion recognition from speech.

First, we investigated a new feature extraction method for ASR using ideal binary masks that are denoted as the goal of computational auditory scene analysis. We built a speaker-independent isolated digit recognition system. We obtained state of the art recognition rate performance in the ideal case with clean speech signals. We compared binary features and MFCCs under noisy conditions and concluded that binary features were more noise robust. One of the important advantages of using binary features would be lower computational power demand due to binary data thus it would be more convenient to use in a small device such as a hearing aid. In addition, we believe that ASR field is a mature area that needs innovative ideas for improvement and that there is much room for advancing our method.

Second, we investigated human top-down auditory attention. We carried out behavioral experiments inspired by Cherry's classic experiments in 1950s, the results of which are still of importance. The participants listened to a monaural mixture of two narratives spoken by the same speaker and reported the words they heard from a list including words from both narratives or words not from either narrative but related to the context of the narratives. We had two sessions, in the first one the listeners were not given any task (they could follow any narrative at any moment), while in the second one, they were asked to follow just one of the narratives. We observed, as expected, a negative correlation between the number of times the words are reported and the temporal and the spectral overlaps of the two narratives in the results of the experiment with no task, while, there was no correlation observed in the results of the experiment with the given task of following just one of the narratives. We concluded that the relation between overlap and speech intelligibility is modulated by the presence of a task. We also designed a computational

top-down attention model as a sequential decision making process in an environment with features missing completely at random, but there is the possibility to gather additional features among the ones that are missing. The saliency of features was given as the difference in confusion (entropy). The proposed top-down mechanism outperformed random saliency model. We also showed that strong task attention model showed less insensitivity to the amount of the confounding overlap, than the weak task attention model as we also observed in the behavioral experiments. We concluded that the top-down attention can undermine the effect of overlapping noise.

Finally, we investigated affect recognition from speech. We designed our own corpus that consisted of short movie clips just with audio and text parts. We combined acoustic and semantic features of speech in valence and arousal dimensions. The results not only showed that fusion of acoustic and semantic features improved the emotion recognition from speech, but also that the effect of acoustic and semantic features differ for different affective dimensions. We concluded that valence was better estimated using semantic features while arousal was better estimated using acoustic features of speech.

Appendix A

Robust Isolated Speech Recognition Using Binary Masks

S.G. Karadogan, J. Larsen, M.S. Pedersen, and J.B. Boldt. Robust isolated speech recognition using binary masks. *Proceedings of European Signal Processing Conference (EUSIPCO)*, pp. 1988-1992, 2010.

ROBUST ISOLATED SPEECH RECOGNITION USING BINARY MASKS

Seliz Gülsen Karadoğan[1], *Jan Larsen*[1], *Michael Syskind Pedersen*[2], *Jesper Bünsow Boldt*[2]

[1] Informatics and Mathematical Modelling, Technical University of Denmark, DK-2800, Kgs. Lyngby, Denmark
[2] Oticon A/S, Kongebakken 9, DK-2765 Smørum, Denmark
{seka, jl}@imm.dtu.dk , {msp,jeb}@oticon.dk

ABSTRACT

In this paper, we represent a new approach for robust speaker independent ASR using binary masks as feature vectors. This method is evaluated on an isolated digit database, TIDIGIT in three noisy environments (car, bottle and cafe noise types taken from the DRCD Sound Effects Library). Discrete Hidden Markov Models are used for the recognition and the observation vectors are quantized with the K-means algorithm using a Hamming distance. It is found that a recognition rate as high as 92% for clean speech is achievable using Ideal Binary Masks (IBM) where we assume prior target and noise information is available. We propose that using a Target Binary Mask (TBM), where only prior target information is needed, performs as good as using IBMs. We also propose a TBM estimation method based on target sound estimation using non-negative sparse coding (NNSC). The recognition results for TBMs with and without the estimation method for noisy conditions are evaluated and compared with those of using Mel Frequency Cepstral Coefficients (MFCC). It is observed that binary mask feature vectors are robust to noisy conditions.

1. INTRODUCTION

Automatic Speech Recognition (ASR) systems have been improving significantly since the 50's. However, there are still many challenges to be surpassed to reach the human performance or beyond. It is well known that one of the key challenges is the robustness under noisy conditions. Another challenge is the need for innovative modeling frameworks. Most of the work has been focusing on the successful representations such as mel frequency cepstral coeffients (MFCC). However, because of a long history of research within the current ASR paradigm, the performance enhancement usually reported is very little. We will suggest a new approach which gives the state of the art performance that is robust to noisy environments.

Since the human auditory system has a great performance, it is tempting to use the human auditory system as an inspiration for an efficient ASR system. Auditory Scene Analysis (ASA) studies perceptual audition and describes the process how the human auditory system organizes sound into meaningful segments [1]. Computational ASA (CASA) makes use of some of the ASA principles and it is claimed that the goal of CASA is the ideal binary mask (*IBM*) [2]. *IBM* is a binary pattern obtained with the comparison of the target and the noise signal energies with prior information of target and noise signals separately. *IBM*s have been shown to improve speech intelligibility when applied to noisy speech signals. The listeners have been exposed to the resynthesized speech signals from the IBM-gated signal and almost perfect recognition results have been obtained even for a signal-to-noise-ratio (*SNR*) as low as -60 dB which corresponds to pure noise [3, 4]. Having proven to make improvements on speech intelligibility of humans, it is inevitable not to make the use of CASA and thus *IBM*s for machine recognition systems. Green et. al. have studied this in [5]. They used CASA as a preprocessor to ASR and used only the time-frequency regions of the noisy speech which are dominated by the target signal to obtain the recognition features. Therefore, they concluded that occluded (incomplete) speech might contain enough information for the recognition.

In this work we go one step further and explore the possibility that not only the occluded speech but the mask itself might carry sufficient information for ASR. The most obvious benefit of this new approach is the simplicity with the use of binary information on the mask. The difficulty about using this method would be the need for the prior information of the target and noise signals to estimate the *IBM*. However, we minimize this need by using Target Binary Mask (*TBM*) where only target information is needed and compared to a speech shaped noise (*SSN*) matching the long term spectrum of a large collection of speakers. Using *TBM*s has also been proven to give high human speech intelligibility [4]. In addition, we propose a *TBM* estimation method based on non-negative sparse coding (NNSC) [6].

This paper will focus on a speaker-independent isolated digit recognizer with hidden Markov models (HMM) using the binary masks as the feature vectors. In Section 2 we give the modeling framework. The experiments and results are explained in Section 3. Finally Section 4 states the conclusion.

2. MODELING FRAMEWORK

2.1 Ideal Binary Masks

The computational goal of CASA, the *IBM*, is obtained by keeping the time-frequency regions of a target sound which have more energy than the interference and discarding the other regions. More specifically, it is one when the target is stronger than the noise for a local criteria (*LC*), and zero elsewhere. The time-frequency (T-F) representation is obtained by using the model of the human cochlea as the basis for data representation [7]. If $T(t,f)$ and $N(t,f)$ denote the target and noise time-frequency magnitude, then the *IBM* is defined as

$$IBM(t,f) = \begin{cases} 1, & \text{if } T(t,f) - N(t,f) > LC \\ 0, & \text{otherwise} \end{cases} \quad (1)$$

Figure 1 shows time-frequency representations of the target, noise and mixture signals. The target is digit six by a male speaker while the noise is *SSN* with 0 dB of *SNR*. The corresponding *IBM* with *LC* of 0 dB is also seen in Figure 1. Calculating an *IBM* requires that the target and the noise are available separately.

LC and *SNR* values in Equation 1 are two important parameters in our system. If *LC* is kept constant, increasing or decreasing the *SNR* makes the mask get closer to all-ones mask or all-zeros mask respectively. The change in *IBM*s for a fixed *LC* with different *SNR* values is shown in Figure 2 for a digit sample. As also seen from this figure, with fixed threshold, low or high *SNR* values result in masks with little or redundant information respectively. Meanwhile, increasing the *SNR* value is identical to decreasing the *LC* value and vice versa. Therefore, the relative criterion $RC = LC - SNR$ was defined in [4] and the effect of *RC* of an *IBM* on speech perception was studied. They calculated *IBM*s with priori target and noise information and multiplied the mixture signal with the corresponding *IBM*s. They,exposed human subjects to resynthesized IBM-gated

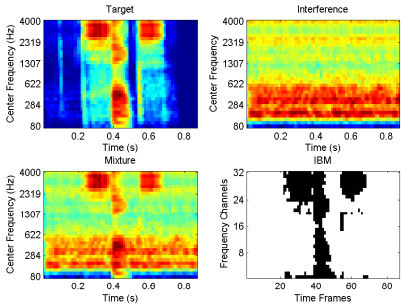


Figure 1: llustration of T-F representations of a target, noise (SSN) and mixture signals with the resultant IBM (0 dB of SNR, 32 frequency channels and window length of 20ms) **red regions:** highest energy, **blue regions:** lowest energy.

mixtures and found high human speech intelligibility (over 95%) for the *RC* range of [-17 dB, 5 dB]. We took this *RC* range as a reference and the results of our ASR system coincided with human speech perception results in terms of *RC* range, which is shown in section 3.

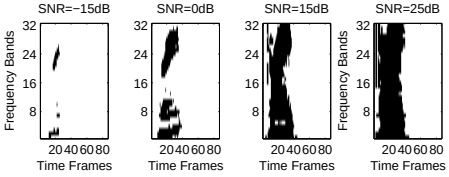


Figure 2: IBMs of digit three with SSN for a fixed LC at 0 dB and for different SNR values .

2.2 Target Binary Masks

The binary mask calculated based on only the target signal was studied and is called Target Binary Mask (*TBM*) [8]. *TBM*s were further investigated in [4] in terms of speech intelligibility and the results were comparable to those of *IBM*s. The definition of *TBM* as seen in equation 2 is very similar to that of *IBM* except that while obtaining *TBM* the target T-F regions are compared to a reference SSN matching the long-term spectrum of the target speaker. (It is also possible to compare the target to a frequency dependent threshold corresponding to the long term spectrum of SSN)

$$TBM(t,f) = \begin{cases} 1, & \text{if } T(t,f) - SSN(t,f) > LC \\ 0, & \text{otherwise} \end{cases} \quad (2)$$

Figure 3 illustrates the T-F representation of a target signal and the mixture signal with cafe noise at 0 dB SNR. That figure also shows the resultant *IBM* and *TBM* patterns with *LC* of 0 dB, and the difference between them is discernible. The *TBM* mimics the target pattern better, whereas the *IBM* pattern depends on the noise type.

Some of the properties of *TBM* can be very practical. First of all, acquiring a *TBM* needs only the priori information of the target. Therefore, estimating the *TBM* can be much more convenient in some applications, especially if speech enhancement techniques are used. In the case of an ASR system that is robust to noise types, use of *TBM*s in the training stage requires less computational effort as opposed to the use of *IBM*s where it is needed to include all *IBM*s for all different noise types in the training stage.

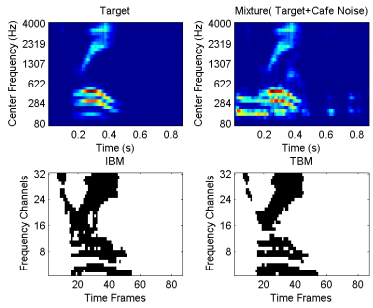


Figure 3: llustration of T-F representations of a target (digit six), mixture (target+cafe noise) and mixture signals with the resultant IBM and TBM **red regions:** highest energy, **blue regions:** lowest energy.

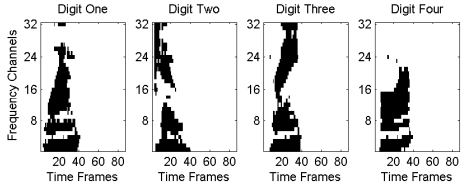


Figure 4: IBMs for different digits for the same speaker

2.3 ASR Using Binary Masks

As mentioned previously, we investigate if the mask itself can be used to recognize different words. The distinctivity of the masks can be observed easily in Figure 4, in which *IBM*s for four different digits with *SNR* of -6 dB using SSN as interference are shown. (Note that *IBM* is identical to *TBM* when the noise type is SSN.) Moreover, as seen in Figure 5 , the masks for different speakers for the same digit are very similar. Thus, the patterns in every mask are characteristic for each digit which results that these patterns are promising representations for speech recognition.

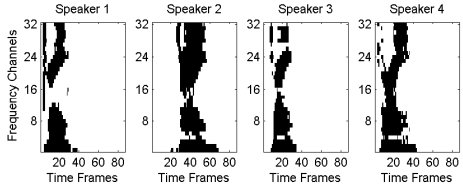


Figure 5: IBMs for digit three for different speakers.

We use a discrete Hidden Markov Model (HMM) as the recognition engine [9]. As the vector quantization method before HMM, we choose to use K-means algorithm, which has been shown to perform as well as many other clustering algorithms and is computationally efficient [10] and proven to be successfully applicable to classify binary data [11]. Figure 6 illustrates the acquisition of the feature vectors to be classified by K-means. We stack the columns of the *IBM* into a vector. The number of columns to be stacked is a parameter that has been optimized for this work (it is 3 for this study) as well as other parameters: the codebook size, the state number of the HMM, the number of frequency bands, and the win-

dow length of the *IBM*. The optimization process can be found in detail in [12].

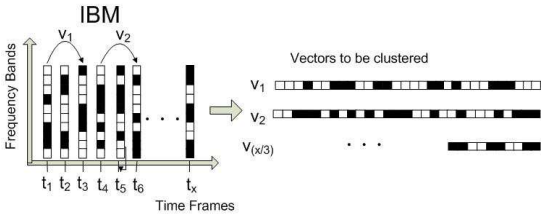


Figure 6: Acquistion of the feature vectors to be clustered by K-means.

The whole system is summarized in Figure 7. First, the masks for training and test data are calculated. The feature vectors obtained from *IBM*s are quantized with K-means to acquire the observed outputs for discrete HMM. One HMM for each digit is trained with the corresponding data. Finally, the test masks are input to each HMM and the test digit is assigned to the one with the highest likelihood. We use only clean data for training. However, for testing we use clean data to see the best performance that can be obtained with our system, an unprocessed mixture signal to see the worst case performances under noisy conditions and finally an estimated target signal from the mixture to see the improved results under noisy conditions.

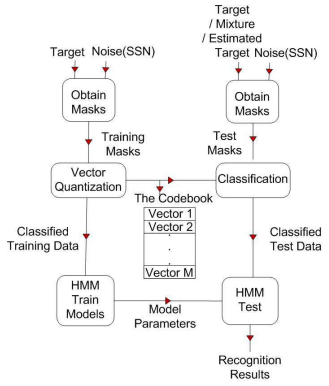


Figure 7: The schematics representation of the system used.

2.4 Estimation of TBMs

Estimation of *TBM* is simpler compared to that of an *IBM* as mentioned previously. Once the target signal is estimated, it is compared to a reference *SSN* signal in the T-F domain. For speech and noise separation, non-negative sparse coding (NNSC), a combination of sparse coding and non-negative matrix factorization, is used [6]. This method was proven to be successful for wind noise reduction in [13], and we took this work as reference for our method.

The principle in NNSC is to factorize the non-negative signal, X into a dictionary W and a code H:

$$X \approx WH. \tag{3}$$

The columns of the dictionary can be considered as the basis and the code matrix can be considered to have the weights for each of the basis vectors constituting the signal X. In our case X is the T-F representation of a signal which is non-negative (details about the acquisition of T-F spectrogram are in section 3). We use the method described in [13] that is based on the algorithm in [14]. W and H are initialized randomly, and updated according to the equations below until convergence:

$$H \longleftarrow H \times \frac{W^T.X}{W^T.W.H+\lambda}, \tag{4}$$

$$W \longleftarrow W \times \frac{X.H^T + W \times (1.(W.H.H^T \times W))}{W.H.H^T + W \times (1.(X.H^T \times W)))}. \tag{5}$$

Here, (.) indicate direct multiplication, while ($\times$) and (_) indicate point wise multiplication and division. 1 is a square matrix of ones of suitable size.

When the speech signal is noisy, and if the noise signal is assumed to be additive, then

$$X = X_s + X_n \approx [W_s W_n] \begin{bmatrix} H_s \\ H_n \end{bmatrix}, \tag{6}$$

where X_s and X_n denote the speech and noise. We precompute the noise dictionary W_n using noise recordings and using equations 4 and 5. We keep this precomputed W_n fixed and learn speech X_s using the following iterative algorithm,

$$H_s \longleftarrow H_s \times \frac{W_s^T.X}{W_s^T.W.H+l_s}, \tag{7}$$

$$H_n \longleftarrow H_n \times \frac{W_n^T.X}{W_n^T.W.H+l_n}, \tag{8}$$

$$W_s \longleftarrow W_s \times \frac{X.H_s^T + W_s \times (1.(W.H.H_s^T \times W_s))}{W.H.H_s^T + W_s \times (1.(X.H_s^T \times W_s)))}, \tag{9}$$

The clean speech is estimated as

$$X_s = W_s H_s. \tag{10}$$

Finally, the *TBM* is estimated by comparing the estimated speech signal X_s to the reference *SSN* signal spectrogram using equation 2. As mentioned previously, different *RC* values lead to masks with different densities and only choosing the right *RC* values leads to high recognition results. However, we learn the right *RC* values for ASR after training and testing with *IBM*s, where we have the pure target and noise signals. (The results can be seen in section 3 in figure 8.) We assume that after NNSC we have the pure target spectrogram. Then, since we also have the reference *SSN* signal spectrogram that is also used during training, we only need to adjust *SNR* and *LC* values for the right *RC* value. However, to obtain the *SNR* between the estimated target and speech, we do not go back to the time domain which would be a waste of time and computational power. Thus, we define a new *SNR* in the T-F domain which is calculated by the ratio between the sum of all T-F bins of the target signal to the sum of all T-F bins of the noise signal and is called as SNR_{TFD}. We observed that $RC_{TFD} = LC_{TFD} - SNR_{TFD}$ range is similar to *RC* range found before (The results can be seen in section 3 in figure 10).

3. EXPERIMENTAL EVALUATIONS

In the experiments, data from TIDIGIT database were used. The spoken utterances of 37 male and 50 female speakers for both training and test data were taken from the database. There are two examples from every speaker for each of 11 digits (zero-nine, oh) making 174 training, 87 test and 87 verification utterances for each digit. The verification set has been used to obtain the optimized parameters for HMM and for NNSC and the final results are obtained using the test set. The experiments were carried out in MATLAB and an HMM toolbox for MATLAB by Kevin Murphy was used [15]. The experiments have also been verified using the HMMs in the Statistical Toolbox of MATLAB. For NNSC the NMF:DTU toolbox

for MATLAB [16] has been adjusted for our system and used. The time-frequency representations of the signals sampled at 8kHz have been obtained using a gammatone filter with 32 frequency channels equally distributed on the ERB scale within the range of [80 Hz, 4000 Hz]. The output from each filterbank channel was divided into 20 ms frames with 10 ms overlap. SSN, car, bottle(the sound of many bottles chinking on a production line) and cafe noise were used through the experiments [17]. A left-to-right HMM with 10 states was used to model each digit. The binary vectors were quantized into a codebook of size 256 with K-means. The HMMs were trained with *IBM*s obtained with *LC* of 0 dB and with different *SNR* values in the range of [-2 dB,16 dB] with 2 dB steps only using *SSN* as the reference noise signal. We compare the method with a standard approach using 20 static MFCC features. MFCC vectors are also stacked as in Figure 6 and all parameters are the same except for the optimized codebook size of 32. The optimal codebook size is smaller since we have less training data for MFCC. One minute of SSN, car, bottle and cafe noise recordings were used to obtain the dictionaries for NNSC. For training, verification or test noise samples different parts of corresponding noise types were used.

Recognition results obtained for the test set for *IBM*s with *SSN* for *LC* of 0 dB and different *SNR* values are presented in Figure 8. The rate curve is bell-shaped, i.e., the rate does not increase monotonously while *SNR* increases. This is because of the previously mentioned fact that either increasing or decreasing the *SNR* value results in masks closer to all-ones or all-zeros masks and thus in the decrease of the recognizability of the masks. Figure 8 shows that 92% recogniton rate is obtained for *RC* of -6 dB. Thus, the mask with *RC* of -6 dB gives the maximum performance.

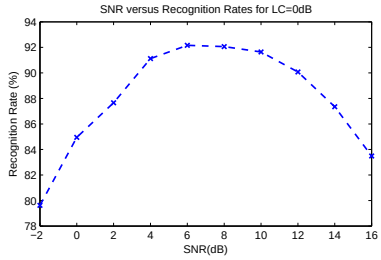


Figure 8: The recognition rates with IBMs for LC=0 dB and SNR=[-2 dB, 16 dB]

If the *LC* value can be adjusted so that the mask is as close to the maximum-performance mask as possible (*RC* is close to -6 dB), we can obtain high recognition results for different *SNR* values. Choosing the correct *LC* value under noisy conditions is a challenge since we know neither the *SNR* value nor the noise spectrogram in real life applications. This problem will be solved by using the NNSC method assuming we have information about the noise characteristics. However, it is reasonable to check the recognition results that can be obtained comparing unprocessed mixture signals to *SSN* with adjusted *LC* values (results are obtained with different *LC* values and the best result is recorded) before exploring that method. Figure 9 shows the recognition rates obtained using HMMs trained with *IBM*s obtained by clean data and *SSN*, with the test set added different noise types at an *SNR* range of [0 dB, 20 dB] (with adjusted *RC* value for the best performance). In that figure, the results obtained using static MFCC features are also shown. It can be seen that using *IBM* features yields more noise-robust recognition rates than using MFCC features. We point out the fact that we used only static MFCC features and did not use any of the improvement methods suggested for MFCC that result in a better performance [18]. Nevertheless, we did not use dynamical features that could be obtained from *IBM*s either. In addition, we believe that the performance of *IBM*s for ASR can also be improved in various ways such as mask estimation methods [19]. Moreover, if we consider the ASR results obtained using MFCC within recent works, our results are comparable [18]. (We cannot make a direct comparison though, since they use a different system and database.) In addition, our method establishes a new route for robust ASR that is open for further improvements. (Some additional results and figures of the whole system can be found at [12]).

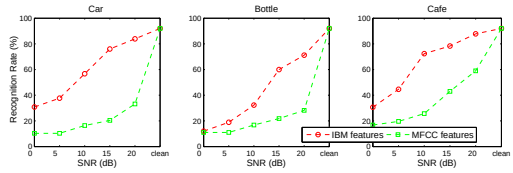


Figure 9: The recognition rates for TBMs and MFCC features at SNR range of [0 dB, 20 dB]

As mentioned previously, for NNSC we needed to the find RC_{TFD} range giving high recognition results. The corresponding results can be seen in Figure 10 and -6 dB of RC_{TFD} gives the maximum performance and RC between -16 dB and 2 dB gives reasonable recognition results (over 80%). The optimized parameters for NNSC for this work are the size of the dictionary of noise and speech, W_n and W_s. Other parameters λ,l_s and ln were just set to be very small numbers taking reference the results in [13]. To find the optimal parameters for the size of W_n and W_s, we checked the recognition results for different size numbers between 4 and 512 for all noise types with SNR_{TFD} of 10 dB and *LC* of 0 dB. We choose 64 for W_n and 128 for W_s based on the results seen in Figure 11.

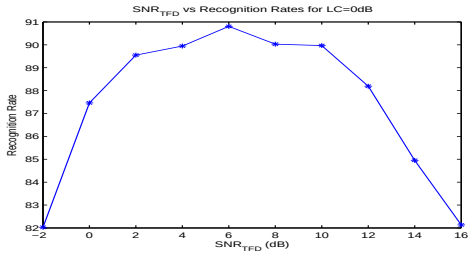


Figure 10: The recognition rates with IBMs for *LC*=0 dB and SNR_{TFD}=[-2 dB, 16 dB]

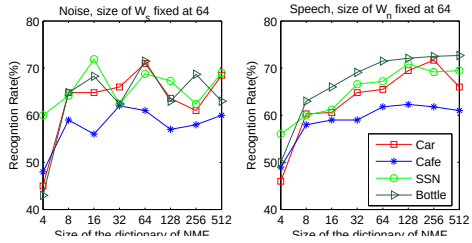


Figure 11: The recognition rates for different size of W_n and W_s

In Figure 12, the recognition rates obtained with noisy mixtures before and after using NNSC is shown (with reference *SSN* at SNR_{TFD} of 0 dB). As seen on the left of this figure, before NNSC,

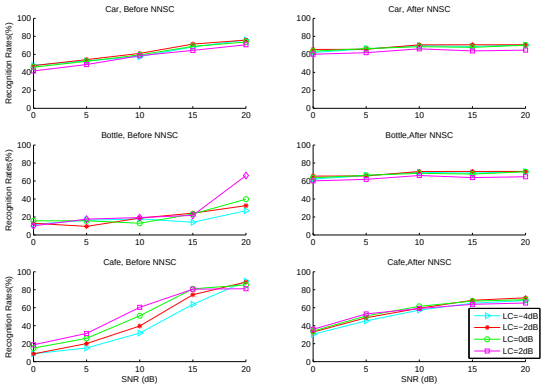


Figure 12: The recognition rates before and after NNSC

different *LC* values within the good RC range found before (-4 dB to 2 dB), result in scattered recognition rates. For cafe noise at 10 dB *SNR*, it is seen that before NNSC the rates can change from 30% to 60% for those different *LC* values. However, after using NNSC to estimate the masks as explained, it is seen that the rates for those *LC* values give the best performances, solving the choice of the right *LC* values for our ASR system. Using NNSC not only solves this problem but also leads to higher recognition results especially for low *SNR* values at the price of a decrease in recognition results for high *SNR* values. However, the decrease in high *SNR* values is not as much as the increase in low ones. Finally, we obtain 60% to 70%, 16% to 73% and 40% to 70% recognition rates for *SNR* values between 0 dB and 20 dB for car, bottle and cafe noises respectively, which are comparable to the state-of-the-art results [18, 20].

4. CONCLUSION

In this paper, we investigated a new feature extraction method for ASR using ideal and target binary masks. It is found that using binary information from the masks directly as feature vectors results in high recognition performance. We constructed a speaker-independent isolated digit recognition system. The experiments were carried out with TIDIGIT database, using discrete HMM as the recognition engine. The K-means algorithm with hamming distance was used for vector quantization. The maximum recognition rate achieved for clean speech is 92%. In addition, the robustness of the binary mask features to different noise types (car,bottle and cafe) was explored and the results were compared to the MFCC features results. A *TBM* estimation method using non-negative sparse coding has been demonstrated to give state of the art performance. It is concluded that noise-robust ASR systems can be built using binary masks.

Acknowledgments: We acknowledge the independent work similar to our work that we became aware of after our model was developed [21].

References

[1] A.S. Bregman, *Auditory Scene Analysis*, Cambridge, MA: MIT Press, 1990.

[2] D. Wang, "On ideal binary mask as the computational goal of auditory scene analysis," *Speech separation by humans and machines*, pp. 181–197, 2005.

[3] D. Wang, U. Kjems, M.S. Pedersen, J.B. Boldt, and T. Lunner, "Speech perception of noise with binary gains," *The Journal of the Acoustical Society of America*, vol. 124, pp. 2303–2307, 2008.

[4] U. Kjems, J.B. Boldt, M.S. Pedersen, T. Lunner, and D. Wang, "Role of mask pattern in intelligibility of ideal binary-masked noisy speech," *The Journal of the Acoustical Society of America*, pp. 1415–1426, 2009.

[5] P.D. Green, M.P. Cooke, and M.D. Crawford, "Auditory scene analysis and hidden Markov model recognition of speech in noise," in *IEEE International Conference on Acoustics Speech and Signal Processing*, 1995, vol. 1, pp. 401–401.

[6] P.O. Hoyer, "Non-negative sparse coding," *Neural Networks for Signal Processing*, pp. 557–565, 2002.

[7] R. Lyon, "A computational model of filtering, detection, and compression in the cochlea," in *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP'82.*, 1982, vol. 7, pp. 1282–1285.

[8] M. C. Anzalone, L. Calandruccio, K. A. Doherty, and L. H. Carney, "Determination of the potential benefit of time-frequency gain manipulation," *Ear Hear*, vol. 27, pp. 480–492, 2006.

[9] L.R. Rabiner, "A tutorial on hidden markov models and selected application in speech recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, 1989.

[10] M. Steinbach, G. Karypis, and V. Kumar, "A comparison of document clustering techniques," in *Text Mining Workshop, in Proc. of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2000)*, 2000, vol. 34, p. 35.

[11] J. Schenk, S. Schwarzler, G. Ruske, and G. Rigoll, "Novel VQ designs for discrete hmm on-line handwritten whiteboard note recognition," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 5096 LNCS, pp. 234–243, 2008.

[12] S.G. Karadogan, J. Larsen, M.S. Pedersen, and J.B. Boldt, "Robust isolated speech recognition using ideal binary masks," http://www2.imm.dtu.dk/pubdb/p.php?5780.

[13] Larsen J. Schmidt, M.N. and Fu-Tien H., "Wind noise reduction using non-negative sparse coding," *IEEE Workshop on Machine Learning for Signal Processing*, pp. 431–436, 2007.

[14] Eggert J. and Körner E., "Sparse coding and nmf," *IEEE International Conference on Neural Networks*, vol. 4, pp. 2529–2533, 2004.

[15] K. Murphy, "Hidden markov model(HMM) toolbox for MATLAB," .

[16] IMM Technical University of Denmark, "Nmf:dtu toolbox," .

[17] The Danish Radio, "The DRCD Sound Effects Library," .

[18] C. Yang, F. K. Soong, and T. Lee, "Static and Dynamic Spectral Features: Their Noise Robustness and Optimal Weights for ASR ," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, no. 3, pp. 1087–1097, 2007.

[19] D. Wang, "Time Frequency Masking for Speech Separation and Its Potential for Hearing Aid Design," in *Trends in Amplification*, 2008, vol. 12, pp. 332–353.

[20] B. Gajic and K.K. Paliwal, "Robust speech recognition in noisy environments based on subband spectral centroid histograms," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 14, no. 2, pp. 600–608, 2006.

[21] A. Narayanan and D.L. Wang, "Robust speech recognition from binary masks," 2010.

Appendix B

How efficient is estimation with missing data?

S.G. Karadogan , L. Marchegiani, L.K. Hansen and J. Larsen. How efficient is estimation with missing data? *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2260-2263, 2011.

HOW EFFICIENT IS ESTIMATION WITH MISSING DATA?

Seliz G. Karadoğan[1] *Letizia Marchegiani*[2] *Lars Kai Hansen*[1] *Jan Larsen*[1]

[1] DTU Informatics, Technical University of Denmark,
DK-2800, Kgs. Lyngby, Denmark
[2] Department of Computer and System Sciences
Sapienza, University of Rome, 00185 Roma, Italy

ABSTRACT

In this paper, we present a new evaluation approach for missing data techniques (MDTs) where the efficiency of those are investigated using listwise deletion method as reference. We experiment on classification problems and calculate misclassification rates (MR) for different missing data percentages (MDP) using a missing completely at random (MCAR) scheme. We compare three MDTs: pairwise deletion (PW), mean imputation (MI) and a maximum likelihood method that we call complete expectation maximization (CEM). We use a synthetic dataset, the Iris dataset and the Pima Indians Diabetes dataset. We train a Gaussian mixture model (GMM). We test the trained GMM for two cases, in which test dataset is missing or complete. The results show that CEM is the most efficient method in both cases while MI is the worst performer of the three. PW and CEM proves to be more stable, in particular for higher MDP values than MI.

Index Terms— Machine learning, supervised learning, missing data techniques

1 Introduction

The reconstruction of degraded audio and video sequences, the analysis of images with missing pixels or occlusions, the manipulation of distorted signals due to sensor failure or outliers are just a few of the wide range of situations in which signal processing faces the missing data problem. In fact, missing data issues are common to most statistical learning domains.

A number of strategies have been investigated for mitigating the missing data problem and many techniques have been proposed. Basically, it is possible to group these techniques in three main categories: *deletion* methods, *imputation* methods and *model-based* methods. In the first, the analysis considers only the present data. The deletion procedure involves removing only the missing input features (*pairwise deletion*) or entire data samples containing them (*listwise deletion*) [1,2]. In the imputation methods, the missing data are replaced with other estimates, so that, like in the pairwise deletion, all the available information is kept and utilized. The simplest way to implement an imputation process is to substitute the missing value of a variable for the mean value of the same variable (*mean imputation*) [3]. In [4], Rubin proposes the concept of *multiple imputation (MI)*, which consists of inserting several values, instead of just one, for each missing instance. This process generates many complete imputed data sets and standard complete data methods are, then, used to examine each of them. The model-based methods, instead, are able to perform directly their analysis on the incomplete set, without changing or ignoring part of the available information. In particular, maximum likelihood (ML) approaches are the most representative in this category and Expectation Maximization algorithms (EM) are often used in this perspective [5], [6].

The behavior of these methods has been explored in literature, depending on the type of missingness scheme, as proposed by Rubin in [7]. Specifically, data are *missing at random (MAR)* if the probability of missingness depends on the observed data, but not on the missing; *missing completely at random (MCAR)* if the probabilities depend on neither the observed and nor the missing data. In the opposite case, data are said to be *missing not at random (MNAR)*. Roth [2] provides a qualitative evaluation of the most common missing data approaches considering scenarios in applied psychology. Allison [1] analyzes advantages and disadvantages of the same methods, on the basis of three criteria: the capability to minimize bias, maximize the use of available information and yield good estimates of uncertainty. Schafer and Graham [8], perform an analysis close to the cited work of Allison using bias and mean square error to evaluate the model estimation accuracy and the behaviour of the standard error to evaluate the margin of the uncertainty. Myrtveit et al. [9] investigate missing data methods in the context of software cost modelling. In particular, the work focuses on the possible benefits that could be obtained thanks to the use of maximum likelihood, multiple imputation and similar response pattern imputation [1] approaches, instead of the listwise deletion one, that is considered the most frequently utilized in their field.

To our knowledge, a standard and general strategy to compare different missing data techniques (MDTs) and to evaluate their performance have not been proposed yet. To fill the gap, we propose a specific definition of efficiency that can be used to analyse how an algorithm operates on missing data. The efficiency of MDTs is computed considering the listwise deletion method as a reference. Specifically, we test the behaviour of the maximum likelihood method in [6] (*Complete EM*), the pairwise deletion and the mean imputation ones in a classification problem, using the Gaussian mixture model [10], with different percentage of missing information in the training set. We calculate the efficiency of an MDT, for different missing data percentages (MDP) where training data is MCAR in two different contexts. In the first one, we use a complete (no missing values) test set, to evaluate how well the model is estimated. In the second one, we use test set with missing values, to evaluate how robust the estimated model is to missing data. We consider the latter as a more realistic scenario. The analysis is performed for synthetic data, the Pima Indians Diabetes and IRIS data sets.

In section 2, we introduce the learning model used and missing data techniques that are evaluated in terms of efficiency, that is defined in section 3. Finally, the experiments and results are discussed in section 4.

[1] identifying the most similar unit without missing information and replacing the missing part with the correspondent values of this input feature

2 Modeling Framework and Methods

2.1 Modeling Framework

The model used within this work is the Gaussian mixture model (GMM) that is used and explained in [10]. Define $\boldsymbol{x}$ as the d-dimensional input feature vector and the associated output, $y \in \{1, 2, \cdots, C\}$, of class labels, assuming C mutually exclusive classes. The joint input/output density is modeled as the Gaussian mixture.

$$p(y, \boldsymbol{x}|\boldsymbol{\theta}) = \sum_{k=1}^{K} P(y|k)p(\boldsymbol{x}|k)P(k) \tag{1}$$

$$p(\boldsymbol{x}|k) = \frac{1}{\sqrt{|2\pi\boldsymbol{\Sigma}_k|}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_k)^\top \boldsymbol{\Sigma}_k^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_k)\right) \tag{2}$$

where K is the number of components, $p(\boldsymbol{x}|k)$ are the component Gaussians mixed with the non-negative priors $P(k)$, $\sum_{k=1}^{K} P(k) = 1$ and the class-cluster posteriors $P(y|k)$, $\sum_{y=1}^{C} P(y|k) = 1$. The k'th Gaussian component is described by the mean vector $\boldsymbol{\mu}_k$ and the covariance matrix $\boldsymbol{\Sigma}_k$. $\boldsymbol{\theta}$ is the vector of all model parameters, i.e., $\boldsymbol{\theta} \equiv \{P(y|k), \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, P(k) : \forall k, y\}$. The joint input/output for each components is assumed to factorize, i.e., $p(y, \boldsymbol{x}|k) = P(y|k)p(\boldsymbol{x}|k)$.

The input density associated with Eq. (1) is given by

$$p(\boldsymbol{x}|\boldsymbol{\theta}_u) = \sum_{y=1}^{C} p(y, \boldsymbol{x}) = \sum_{k=1}^{K} p(\boldsymbol{x}|k)P(k),$$

where $\boldsymbol{\theta}_u \equiv \{\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, P(k) : \forall k, y\}$. Assuming a 0/1 loss function the optimal Bayes classification rule is $\widehat{y} = \max_y P(y|\boldsymbol{x})$ where[2]

$$P(y|\boldsymbol{x}) = \frac{p(y, \boldsymbol{x})}{p(\boldsymbol{x})} = \sum_{k=1}^{K} P(y|k)P(k|\boldsymbol{x})$$

with $P(k|\boldsymbol{x}) = p(\boldsymbol{x}|k)P(k)/p(\boldsymbol{x})$.

Define data set of labeled examples $\mathcal{D}_l = \{\boldsymbol{x}_n, y_n; n = 1, 2, \cdots, N_l\}$. The negative log-likelihood for the data sets, which are assumed to consist of independent examples, is given by

$$L = -\log p(\mathcal{D}|\boldsymbol{\theta}) = -\sum_{n\in\mathcal{D}_l} \log \sum_{k=1}^{K} P(y_n|k)p(\boldsymbol{x}_n|k)P(k)$$

The model parameters are estimated with an iterative modified EM algorithm [11]:

1. To initialize the mean ($\boldsymbol{\mu}_0$) and covariance ($\boldsymbol{\Sigma}_0$) matrices, all train data set is considered as one normal distribution. In the case of missing data, the calculations are done using only observed data and the $\boldsymbol{\Sigma}_0$ is regularized (see section 2.2). Then, since random points from the distribution can not be taken as cluster center points because of missing data, we draw L random samples using the $\boldsymbol{\mu}_0$ and $\boldsymbol{\Sigma}_0$, and get rid of outliers. Instead of taking random center points from the remaining samples, we use KKZ method assuming the clusters will be distant from each other [12]. The KKZ method is as the following:
   - The first center point is taken as the sample having the largest L_2 norm
   - Other center points are calculated as having the largest distance to the closest center points

[2] The dependence on $\boldsymbol{\theta}$ is omitted.

2. Compute posterior component probability for all $n \in \mathcal{D}_l$:

$$p(k|y_n, \boldsymbol{x}_n) = \frac{P(y_n|k)p(\boldsymbol{x}_n|k)P(k)}{\sum_k P(y_n|k)p(\boldsymbol{x}_n|k)P(k)}. \tag{3}$$

3. For all k, update mean vectors and covariance matrices

$$\boldsymbol{\mu}_k = \frac{\sum_{n\in\mathcal{D}_l} \boldsymbol{x}_n P(k|y_n, \boldsymbol{x}_n)}{\sum_{n\in\mathcal{D}_l} P(k|y_n, \boldsymbol{x}_n)}, \quad \boldsymbol{\Sigma}_k = \frac{\sum_{n\in\mathcal{D}_l} \boldsymbol{S}_{kn} P(k|y_n, \boldsymbol{x}_n)}{\sum_{n\in\mathcal{D}_l} P(k|y_n, \boldsymbol{x}_n)}$$

where $\boldsymbol{S}_{kn} = (\boldsymbol{x}_n - \boldsymbol{\mu}_k)(\boldsymbol{x}_n - \boldsymbol{\mu}_k)^\top$.

4. For all k update cluster priors and class cluster posteriors

$$P(k) = \frac{\sum_{n\in\mathcal{D}_l} P(k|y_n, \boldsymbol{x}_n)}{N_l}, \quad P(y|k) = \frac{\sum_{n\in\mathcal{D}_l} \delta_{y_n} P(k|y_n, \boldsymbol{x}_n)}{\sum_{n\in\mathcal{D}_l} P(k|y_n, \boldsymbol{x}_n)}$$

2.2 Pairwise Deletion

In pairwise (PW) method, the only difference made on the model we use, is the update of posterior input density $p(\boldsymbol{x}_n|k)$, the mean vector μ_k and covariance matrix $\boldsymbol{\Sigma}_k$. To update those, observed data for each variable or pair of variables are used. However, the estimated covariance matrix is unbiased and is not guaranteed to be positive semi definite. We regularize the covariance matrix by inflating the diagonal elements by the factor $(1 + h)$ as in Eq. (4) which is commonly used approach [13] given by

$$\boldsymbol{\Sigma}^{'} = \boldsymbol{\Sigma} + h\boldsymbol{I} \tag{4}$$

where $\boldsymbol{I}$ is the identity matrix and h is a regularization parameter. h is determined in the following way:

$$\boldsymbol{\Sigma}^{'} = \boldsymbol{\Sigma} + h\boldsymbol{I} = \boldsymbol{V}\boldsymbol{U}\boldsymbol{V}^{-1} + h\boldsymbol{V}\boldsymbol{V}^{-1} = \boldsymbol{V}(\boldsymbol{h}\boldsymbol{I} + \boldsymbol{U})\boldsymbol{V}^{-1} \tag{5}$$

where $\boldsymbol{V}\boldsymbol{U}\boldsymbol{V}^{-1}$ is the eigenvalue decomposition of the covariance matrix $\boldsymbol{\Sigma}$, where $\boldsymbol{V}$ is the square matrix whose i'th column is the eigenvector q_i of $\boldsymbol{\Sigma}$ and $\boldsymbol{U}$ is the diagonal matrix whose diagonal elements are the corresponding eigenvalues. Then, we choose h such that $(h + U) > 0$ to have nonnegative eigenvalues in regularized covariance matrix.

2.3 Mean Imputation

Mean imputation (MI) method is a replacement technique where a missing variable is replaced by the corresponding mean value [3]. The model we use is not effected in this method, since we have complete data after imputation. This method keeps all data, and is easy to implement. However, the variance estimates are lessened as more means are added.

2.4 Complete Expectation Maximization

This method is a maximum likelihood missing data technique that is proposed in [6]. EM is used both for the estimation of model components and for dealing with missing data. Posterior component probability, $p(k|y_n, \boldsymbol{x}_n)$ is again calculated as in Eq. (3), but only on observed dimensions. To update the mean vector, $E[\boldsymbol{x}_n^m|\boldsymbol{x}_n^o]$ is substituted for missing components of $\boldsymbol{x}_n$, and to update the covariance matrix, $E[\boldsymbol{x}_n^m \boldsymbol{x}_n^{m^T}|\boldsymbol{x}_n^o]$ is substituted for outer product matrices containing missing components:

$$E[\boldsymbol{x}_n^m|\boldsymbol{x}_n^o] = \mu_n^m + \boldsymbol{\Sigma}_n^{mo}\boldsymbol{\Sigma}_n^{oo^{-1}}(\boldsymbol{x}_n^o - \mu_n^o),$$

$$E[\boldsymbol{x}_n^m \boldsymbol{x}_n^{m^\top}|\boldsymbol{x}_n^o] = \boldsymbol{\Sigma}_n^{mm} - \boldsymbol{\Sigma}_n^{mo}\boldsymbol{\Sigma}_n^{oo^{-1}}\boldsymbol{\Sigma}_n^{mo^T} + E[\boldsymbol{x}_n^m|x_n^o]E[\boldsymbol{x}_n^m|\boldsymbol{x}_n^o]^\top.$$

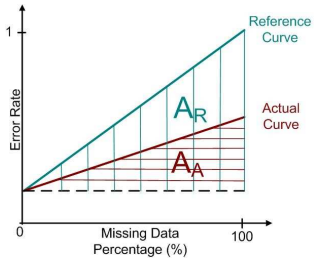


Fig. 1: The illustration for the efficiency calculation method used.

3 Efficiency Definition

As data have more missing values, the resultant error rate (ER) gets higher due to lack of information. However, the resultant MDP-ER curve is different for different missing data techniques (MDTs). In this work, we use the curve for listwise deletion (LW) method as the reference. In other words, we calculate how efficient a technique makes use of data with missing values instead of simply ignoring them. As seen in Figure 1, the definition of efficiency (Eff) is obtained by calculating the area under the reference and actual curves (curves of MDTs investigated) as in Eq. (6). When the actual curve is the same as the reference curve, the efficiency is 0%, while it is 100%, when it is a horizontal line (i.e. ER is not effected as MDP changes, the method is completely robust to MDP).

$$\text{Eff } \% = \frac{A_R - A_A}{A_R} \times 100 \tag{6}$$

4 Experimental Evaluations

The experiments are carried out using MATLAB on synthetically generated data and two datasets from the UCI archive, Iris and Pima-Indian-Diabetes [14]. MDP is determined randomly (MCAR). The experiment is done such that not all values can be missing in one observation (if all data in all directions are missing it would be equal to deleting it, so reducing training data as in our reference method). We experiment how the misclassification rate (MR) changes with MDP and calculate the efficiency (Eq. (6)) using those results for different MDP values. We carry out experiments in two cases. In Case 1, the test dataset also have missing values with same MDP as for training. In Case 2 the test data are complete. Case 2 investigates how well the model is estimated, while the case 1 how robust the estimated model is to missing data. We make 100 iterations for each experiment, while changing MDP between 0% and 70%.

4.1 Synthetic Dataset

The algorithm is tested on synthetic data. The multidimensional input data is generated by a Gaussian mixture model. The number of mixtures K, is 3. The difficulty of the problem is determined using the SNR calculation:
Let ds_{kl} be the distance between μ_k and μ_l, eig_k be a vector consisting of eigenvalues of Σ_k and mean() be the arithmetic mean operator, then

$$\text{SNR}_{\text{dB}} = 10 \log \left(\frac{(\text{mean}(\sum_{1 \le k \le K, k \le l \le K} \text{ds}_{\text{kl}}))^2}{\text{mean}(\sum_{1 \le k \le K} \text{mean}(\text{eig}_k))} \right)$$

We use SNR of 10 dB, for a 10 dimensional data. 150 observations are generated for both training and test sets. Figure 2 shows the first three principal components plotted against each other for data used for this work.

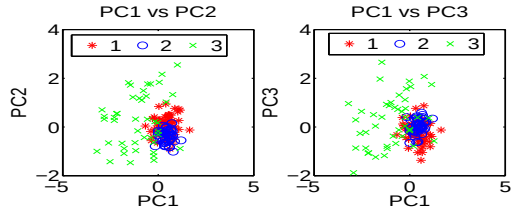


Fig. 2: The principal components (PCs) plot for the data generated with 3 different classes.

The Figure 3 shows the results for synthetic data generated. In case 1, CEM is the most efficient method, however, PW is competitive to it. CEM gives an efficiency of 40%, even at MDP of 70%. MI is clearly the worst method in terms of efficiency. The efficiency of MI decreases as MDP gets higher, while CEM and PW give more stable efficiency results. In case 2, results are similar and still CEM is the best. While MI performs better compared to case 1, CEM is slightly worse.

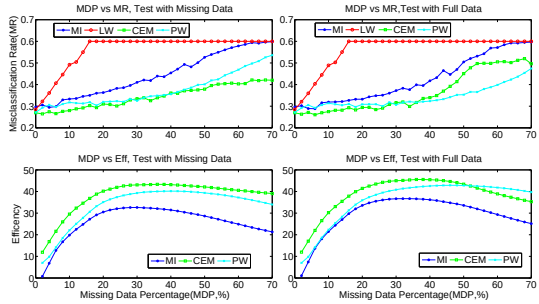


Fig. 3: The results for synthetically generated data. **Left**: Test set with missing (incomplete) data. **Right**: Test set with full (complete) data. **Top**: MR plot against MDP. **Bottom**: Eff plot against MR.

4.2 Iris Dataset

Iris dataset is one of the most commonly used datasets in machine learning literature. It consists of 3 classes of 50 instances each referring to a type of iris plant with 4 attributes. One class is linearly separable from the others; the other two are not linearly separable from each other. We use 100 instances for train and 50 instances for test sets.
We show the results for this dataset in Figure 4. In case 1, CEM is still the most efficient method, MI and PW show a similar behaviour. CEM gives an efficiency of 70%, even at MDP of 70%. In case 2, PW is worse than MI and CEM is still the best method. Compared to case 1, the efficiency of CEM and PW is lower while the efficiency of MI is higher.

4.3 Pima Indians Diabetes Dataset

Pima Indians Diabetes Dataset contains 2 classes that are diabetes positive or negative with 7 attributes (age, pregnancy number, etc.). We use 200 instances for train and 200 instances for test sets.
The results are shown in Figure 5. Both in case 1 and case 2, CEM overcomes other two methods, whereas PW and MI give similar re-

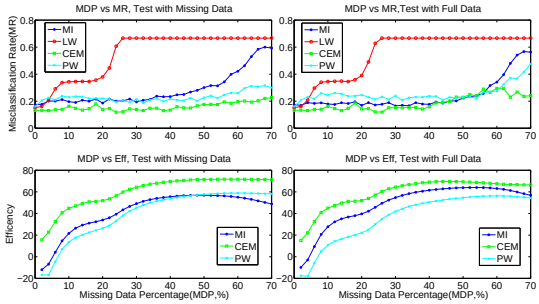


Fig. 4: The results for Iris dataset. **Left**: Test set with missing (incomplete) data. **Right**: Test set with full (complete) data. **Top**: MR plot against MDP. **Bottom**: Eff plot against MR.

sults. The efficiency of CEM at MDP of 70% is around 20%, not as high as other datasets, but still giving the highest performance.

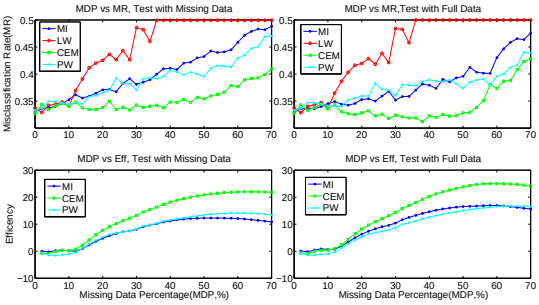


Fig. 5: The results for Pima Indians Diabetes dataset. **Left**:Test set with missing (incomplete) data. **Right**: Test set with full (complete) data. **Top**: MR plot against MDP. **Bottom**: Eff plot against MR.

4.4 General Discussion

We observe that, generally CEM is the most efficient missing data method, while PW is worse than CEM, but still slightly better than MI especially for high MDP values. The results coincide with previous work [1, 15]. In [1], where they compare missing data methods using different criteria (the capability to minimize bias, maximize the use of available information and yield good estimates of uncertainty), ML methods are found to be the best. In [15], where they compare 6 different methods including PW and EM methods, the results again support ML approaches. Although CEM and PW perform well for both cases we experimented, we observe that they are more efficient to use when test data set also has missing values. MI is more efficient to use when we have a complete test data set. Thus, MI is better at estimating the model, but the estimated model is not that robust to missing data in test set, and vice versa for CEM. Another observation made from the results is that CEM and PW give more stable results for higher MDP values, so it would be more trustworthy to use them in situations where MDP for test set is undetermined. Although MI turned out to be the least efficient approach, it would be still acceptable to use it especially for low MDP values, since it is very easy to implement and clearly computationally less expensive.

5 Conclusion

We proposed a new evaluation approach for missing data techniques (MDTs) where the efficiency of those are investigated using listwise deletion method as reference. We experimented on classification problems and calculated misclassification rate (MR) for different missing data percentages (MDPs). We compared three different MDTs: pairwise deletion (PW), mean imputation (MI) and complete expectation-maximization (CEM). We used synthetic dataset, Iris dataset and Pima Indians Diabetes dataset. We used a Gaussian mixture model trained with missing completely at random data. We tested for missing or complete dataset. The results showed that CEM was the most efficient method in both cases while MI was the worst of the three. We observed that PW and CEM are more stable with respect to especially higher MDP values than MI. We also observed that MI performed better with complete test set, so was better at estimating the model, but the estimated model was not that robust to missing data in test set, vice versa for PW and CEM.

6 References

[1] P.D. Allison, *Missing Data*, Quantitative Applications in the Social Sciences. Sage Publications, 2001.

[2] P.L. Roth, "Missing data: A conceptual review for applied psychologists," *Personnel Psychology*, vol. 47, no. 3, pp. 537–560, 1994.

[3] A.R.T. Donders, G.J.M.G. van der Heijden, T. Stijnen, and K.G.M. Moons, "Review: a gentle introduction to imputation of missing values," *Journal of Clinical Epidemiology*, vol. 59, no. 10, pp. 1087–1091, 2006.

[4] D.B. Rubin, *Multiple Imputations for nonresponse in surveys*, New York: Wiley, 1987.

[5] A.P. Dempster, N.M. Laird, and D.B.Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society*, vol. 39, no. 1, pp. 1–38, 1977.

[6] Z. Ghahramani and M.I. Jordan, "Supervised learning from incomplete data via an EM approach," in *Advances in Neural Information Processing Systems 6*. 1994, pp. 120–127, Morgan Kaufmann.

[7] D.B. Rubin, "Inference and missing data," *Biometrika*, vol. 63, no. 3, pp. 581–592, 1976.

[8] J.L. Schafer and J.W. Graham, "Missing data: Our view of the state of the art," *Psychological methods*, vol. 7, no. 2, pp. 147–177, 2002.

[9] I. Myrtveit, E. Stensrud, and U.H. Olsson, "Analyzing data sets with missing data: An empirical evaluation of imputation methods and likelihood-based methods," *IEEE Transactions on Software Engineering*, vol. 27, no. 11, pp. 999–1013, 2001.

[10] J. Larsen, A. Szymkowiak, and L.K. Hansen, "Probabilistic hierarchical clustering with labeled and unlabeled data," *International Journal of Knowledge Based Intelligent Engineering Systems*, vol. 6, no. 1, pp. 56–63, 2002.

[11] L.K. Hansen, S. Sigurdsson, T. Kolenda, F.A. Nielsen, U. Kjems, and J. Larsen, "Modeling text with generalizable Gaussian mixtures," in *ICASSP'00. Proceedings*. IEEE, 2000, vol. 6, pp. 3494–3497.

[12] T. Su and J.G. Dy, "In search of deterministic methods for initializing K-means and Gaussian mixture clustering," *Intelligent Data Analysis*, vol. 11, no. 4, pp. 319–338, 2007.

[13] T. Schneider, "Analysis of incomplete climate data: Estimation of mean values and covariance matrices and imputation of missing values," *Journal of Climate*, vol. 14, no. 5, pp. 853–871, 2001.

[14] UC Irvine Machine Learning Repository, "http://archive.ics.uci.edu/ml/," .

[15] D.A. Newman, "Longitudinal modeling with randomly and systematically missing data: A simulation of ad hoc, maximum likelihood, and multiple imputation techniques," *Organizational Research Methods*, vol. 6, no. 3, pp. 328, 2003.

Appendix C

What to measure next to improve decision making? On top-down task driven feature saliency

---

L.K. Hansen, S.G. Karadogan and L. Marchegiani. What to measure next to improve decision making? On top-down task driven feature saliency. *IEEE Symposium Series on Computational Intelligence (SSCI)*, pp. 1-7, 2011.

What to measure next to improve decision making? On top-down task driven feature saliency

Lars Kai Hansen, Seliz Karadogan, and Letizia Marchegiani[1]
DTU Informatics, Technical University of Denmark
DK-2800 Kgs. Lyngby, Denmark
lkh,seka,malet@imm.dtu.dk
[1]Permanent address: Department of Computer and System Sciences,
Sapienza, University of Rome,
00185 Roma, Italy

***Abstract*—Top-down attention is modeled as decision making based on incomplete information. We consider decisions made in a sequential measurement situation where initially only an incomplete input feature vector is available, however, where we are given the possibility to acquire additional input values among the missing features. The procecure thus poses the question *what to do next?* We take an information theoretical approach implemented for generality in a generative mixture model. The framework allows us reduce the decision about what to measure next in a classification problem to the estimation of a few one-dimensional integrals per missing feature. We demonstrate the viability of the framework on four well-known classification problems.**

I. Introduction

We are interested in decisions made in the face of incomplete measurements as a model of top-down task driven attention. Thus, we consider top-down attention to be a situation in which vague information about a decision is presented and our brain has to infer which additional 'measurement' to perform.

A related two-stage mechanism for visual attention has been proposed by Torralba et al. see e.g., [1]. In their work there is no explicit model of the 'task' as in the present work, but rather a global path that serves to guide focal attention. The incomplete measurement situation is closely related also to the missing data classification problem, see e.g., [2].

Thus, we are interested in the situation in which we are given an initial incomplete measurement and there is the possibility to obtain more feature values, hence, it is important to make decisions about which feature to probe next. Such a decision will be based on a trained classifier model and the particular values of initial incomplete measurement. We refer to this situation as the *sequential measurement problem*. While not as well studied as the basic missing data problem, the sequential measurement problem has been studied in several application areas including active vision [3], [4], petrophysics and medical diagnosis [5], geophysics [6], and robot navigation [7]. These previous efforts all use an information theoretic approach but implemented in specialized domain dependent representations, hence, less general than the present approach.

The main novelty of the present contribution is to demonstrate the computational feasibility of a general strategy for solving the sequential measurement problem if implemented in a *generic* Gaussian Discrete mixture model for signal detection. Our approach is statistical modeling and we aim to solve the task using machine learning methods. In particular we distill the solution to a simple two-stage process. First, we train a generative classification model based on complete training data. Secondly, we apply the trained model to test data potentially with missing features. In the case of a test datum with missing features we compute the input feature saliency of the missing features conditional on the available features. The saliency is given as the expected reduction in entropy of the classification model over the output probabilities. To test the quality of the top-down saliency estimator we classify the test data with and without the additional feature, as well as with a randomly selected additional feature.

The sequential measurement problem is a special case of experimental design. The preferred information theoretic approach was devised in 1956 by Lindley [8] and has been applied and rediscovered several times in specific applications. Under the information theoretic approach the experiment is designed so as to reduce the entropy of a target distribution, i.e., the most informative experiment. The complexity of estimating the information has a crucial dependence on the representation of the signal detection problem. In this contribution we implement the information theoretic approach within a Gaussian mixture model in which both marginalization and conditioning are of relatively low complexity [2].

The paper is organized as follows, in the following section we introduce the information theoretical measure of task driven saliency, and we present a Gaussian Discrete generative classifier model. In Section III we present the results of four different classification experiments in which we can simulate the missing feature problem and the viability of the top-down saliency as a solution to the sequential measurement problem. Section IV contains concluding remarks.

II. Information theoretical top-down saliency

While training and classification with missing input features are well understood, see e.g., [2], the sequential measurement problem is less studied. We follow Kappen et al. [3] and consider the sequential measurement process as a two-stage

process. We represent the signal detection problem by probability distribution over a set of C signal groups or classes indexed by the discrete variable c, $(c = 1, ..., C)$. Initially we have access to a vectorial observation $\boldsymbol{x}$ with components x_i, $i = 1, ..., I$. The second step concerns an additional measurement z_j which is obtained by attending to a specific channel j, chosen among the set of missing features $\boldsymbol{z}$ with components $z_1, ..., z_J$.

Let the joint probability of the classes and all features observed and missing be denoted $p(c, \boldsymbol{x}, \boldsymbol{z})$. The goal is to make decisions about c. Here we will aim to minimize the error rate, hence we invoke Bayesian decision theory and choose c according to the posterior distribution $p(c|...)$. The condition depends on the stage in the sequential measurement process. Initially, the information available is $\boldsymbol{x}$, thus the relevant probability is [2]

$$\begin{aligned} p(c|\boldsymbol{x}) &= \int p(c, \boldsymbol{z}|\boldsymbol{x}) d\boldsymbol{z} \\ &= \frac{\int p(c, \boldsymbol{x}, \boldsymbol{z}) d\boldsymbol{z}}{\sum_{c=1}^{C} \int p(c, \boldsymbol{x}, \boldsymbol{z}) d\boldsymbol{z}}. \end{aligned} \quad (1)$$

Using the top down attention mechanism we will select an additional feature $\boldsymbol{z}_j$, which will result in the distribution

$$\begin{aligned} p(c|\boldsymbol{x}, z_j) &= \sum_{c=1}^{C} \int p(c, \boldsymbol{z}|\boldsymbol{x}) \prod_{i \neq j} dz_i \\ &= \frac{\int p(c, \boldsymbol{x}, \boldsymbol{z}) \prod_{i \neq j} dz_i}{\sum_{c=1}^{C} \int p(c, \boldsymbol{x}, \boldsymbol{z}) \prod_{i \neq j} dz_i} \end{aligned} \quad (2)$$

The information value of this choice is given as the difference in confusion (entropy) before and after the second measurement, which will depend on the particular outcome of the sequential measurement, z_j,

$$\begin{aligned} \Delta S_j(\boldsymbol{x}, z_j) &= \sum_{c=1}^{C} \log p(c|\boldsymbol{x}, z_j) p(c|\boldsymbol{x}, z_j) \\ &- \sum_{c=1}^{C} \int \log p(c, \boldsymbol{z}|\boldsymbol{x}) p(c, \boldsymbol{z}|\boldsymbol{x}) d\boldsymbol{z} \end{aligned} \quad (3)$$

As z_j is unknown at this stage in the process we are forced to average $\Delta S_j(\boldsymbol{x}, z_j)$ with respect to this variable given the information we have access to, i.e., with respect to the distribution of z_j conditioned on the initial measurement $\boldsymbol{x}$. This procedure provides us with the *expected information gain* of measuring the value of feature j,

$$\begin{aligned} G_j(\boldsymbol{x}) &\equiv \int \Delta S_j(\boldsymbol{x}, z_j) p(z_j|\boldsymbol{x}) dz_j \\ &= \sum_{c=1}^{C} \int \log p(c|\boldsymbol{x}, z_j) p(c, z_j|\boldsymbol{x}) dz_j \\ &\quad - \sum_{c=1}^{C} \int \log p(c, \boldsymbol{z}|\boldsymbol{x}) p(c, \boldsymbol{z}|\boldsymbol{x}) d\boldsymbol{z}. \end{aligned} \quad (4)$$

The information gain can be used to rank features in importance [8], [3]. Note, the second term does not depend on j, hence, can be neglected in the saliency estimate.

A. Gaussian Discrete mixture model

The Gaussian Discrete mixture model (GDMM) is a generative model of the joint distribution, see e.g., [9],

$$p(c, \boldsymbol{x}, \boldsymbol{z}) = \sum_{k=1}^{K} p(k) p(c|k) p(\boldsymbol{x}, \boldsymbol{z}|k) \quad (5)$$

where K is the number of components, $p(k)$ are component probabilities, $p(c|k)$ is a $C \times K$ probability table, and $p(\boldsymbol{x}, \boldsymbol{z}|k)$ are K Gaussian pdfs. We choose a generative representation to allow for modeling of input dependencies which is necessary in order to make inference about missing features. Maximum likelihood parameter estimation in the GDMM leads to a straightforward generalization of expectation maximization algorithm for conventional mixtures.

B. Information gain in the GDMM

Introducing the generative model Eq. (5) in the information gain and using $p(k|\boldsymbol{x}) = p(k)p(\boldsymbol{x}|k)/p(\boldsymbol{x})$, we obtain

$$\begin{aligned} G_j(\boldsymbol{x}) &= \sum_{c=1}^{C} \sum_{k=1}^{K} p(c|k) p(k|\boldsymbol{x}) \times \\ &\qquad \int \log \left[p(c, \boldsymbol{x}, z_j) \right] p(z_j|\boldsymbol{x}, k) dz_j \\ &- \sum_{k=1}^{K} p(k|\boldsymbol{x}) \int \log \left[p(\boldsymbol{x}, z_j) \right] p(z_j|\boldsymbol{x}, k) dz_j \\ &+ \text{const.} \end{aligned} \quad (6)$$

where $p(c, \boldsymbol{x}, z_j) = \sum_{k=1}^{K} p(k) p(c|k) p(\boldsymbol{x}, z_j|k)$ and $p(\boldsymbol{x}, z_j) = \sum_{k=1}^{K} p(k) p(\boldsymbol{x}, z_j|k)$. Thus, computing G for all I features amounts to computing $Q = I * (C + 1) * K$ one-dimensional integrals over Gaussian measures $p(z_j|\boldsymbol{x}, k) = \mathcal{N}(\mu_j(\boldsymbol{x}, k), \sigma_j^2(\boldsymbol{x}, k))$ with

$$\begin{aligned} \mu_j(\boldsymbol{x}, k) &= \mu_{j,k} - \boldsymbol{\Sigma}_{z_j, \boldsymbol{x}, k} \boldsymbol{\Sigma}_{\boldsymbol{x}, \boldsymbol{x}, k}^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_{\boldsymbol{x}, k}) \\ \sigma_j^2(\boldsymbol{x}, k) &= \sigma_{j,k}^2 - \boldsymbol{\Sigma}_{z_j, \boldsymbol{x}, k} \boldsymbol{\Sigma}_{\boldsymbol{x}, \boldsymbol{x}, k}^{-1} \boldsymbol{\Sigma}_{\boldsymbol{x}, z_j, k}. \end{aligned} \quad (7)$$

In these expressions $\mu_{j,k}, \sigma_{j,k}^2$ are the mean and variance of the jth feature in the kth component, while $\boldsymbol{\Sigma}_{a,b,k}$ is the part of the covariance matrix of the k component corresponding to variable sets a, b.

III. Experimental Evaluation

A. Design of experiments

The main objective of the proposed sequential measurement strategy is to reduce the error rate of the ensuing decision classification problem. We therefore analyze four well-known benchmark problems the UCI depository [10], to test how efficient the saliency estimate is relative to simpler alternatives. The four data sets we have chosen for illustration all have heterogeneous input feature sets.

We first train a Gaussian Discrete mixture model on a training set on N_{train} data points. On the remaining N_{test} data points we simulate an incomplete measurement situation in which only $D_0 < D$ features are available for the estimation

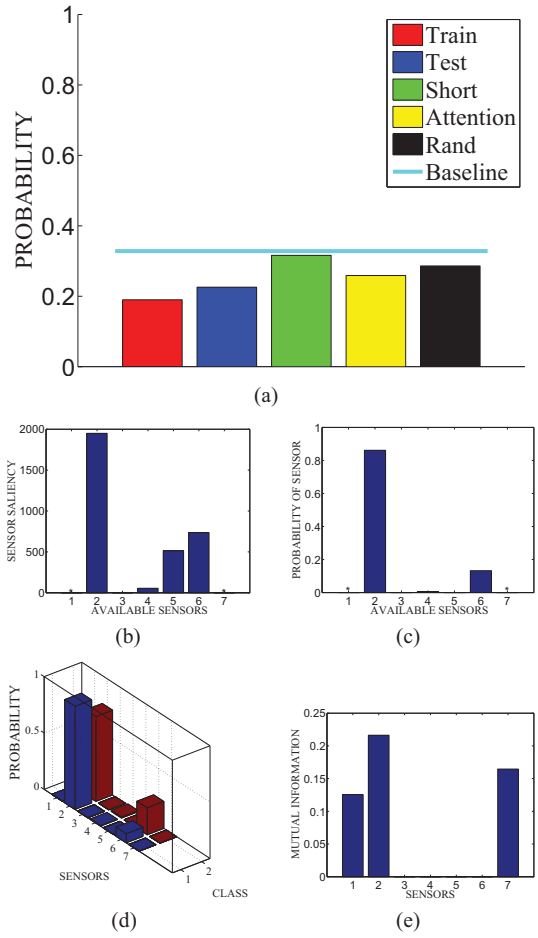


Fig. 1. Pima indian diabetes diagnoses problem. Seven input features are considered. Here we simulate incomplete measurement, in which features $(1, 7) = (\#\text{pregnancies}, \text{age})$ are given. The panels show (a) Error rates in the training set ($N_{\text{train}} = 200$) and the test set ($N_{\text{test}} = 332$) for complete data, test error using only the initial feature set $(1, 7)$, and test set using $(1, 7)$, and the feature chosen among $2 - 6$ by the top down saliency estimate, and finally the test error obtained using features (1,7) and a randomly chosen additional feature; (b) Estimated information saliency obtained on the test data, given the incomplete feature vector (1,7); (c) Frequency of selection of features 2-6; (d) Frequency of selection in test cases within the two classes; (e) The $\log_2$ mutual information between features and class label.

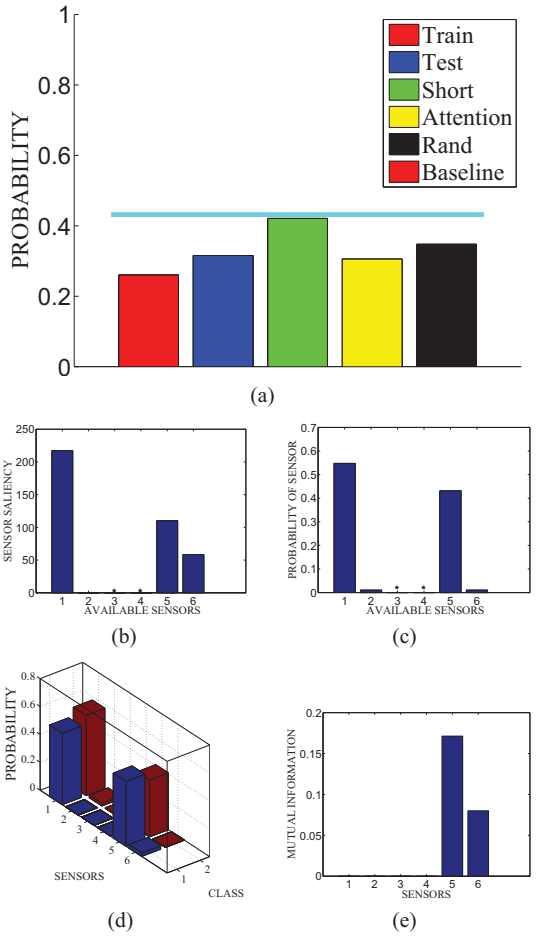


Fig. 2. Liver disorder problem. Seven input features are considered. Here we simulate incomplete measurement, in which features $(3, 4)$ only are provided. (a) Error rates in the training set of complete data ($N_{\text{train}} = 200$) and the test set ($N_{\text{test}} = 95$) using complete data, the test error using the initial feature set $(3, 4)$, and the test error using $(3, 4)$ and the feature chosen by the top down saliency estimate, and finally the test error using features (3,4) and a randomly chosen additional feature; (b) Estimated information saliency obtained on the test data, given the incomplete feature vector (3,4); (c) Frequency of selection of the additional features; (d) Frequency of selection of features in test cases within the two classes; (e) The $\log_2$ mutual information between features and class label.

of saliency, where D is the number of features in the classification problem. We test the performance of two different strategies, 1) choosing the feature with the highest saliency, and 2) choosing a random feature and compare with a classifier that has access to all features and one that only has access to the original two features of the incomplete measurement.

The Gaussian Discrete model is trained with a variable number of components (K). Component covariance matrices were estimated with a simple Wishart prior with a diagonal mean covariance matrix of unit variance. Prior to training all variables were normalized to zero mean and unit variance. The initial training phase was carried out using an EM procedure, which is straightforward to generalize to incorporate the table structure of the Gaussian Discrete model. The training phase involved a multi-start procedure with 10 random initializations and further 1000 EM iterations were carried out on

the initialization that lead to the lowest training error rate in classification. Classification was done according to the posterior distribution

$$p(c|\boldsymbol{x},\boldsymbol{z}) = \frac{\sum_{k=1}^{K} p(k)p(c|k)p(\boldsymbol{x},\boldsymbol{z}|k)}{\sum_{k=1}^{K} p(k)p(\boldsymbol{x},\boldsymbol{z}|k)} = \sum_{k=1}^{K} p(k)p(c|k)p(k|\boldsymbol{x},\boldsymbol{z}) \quad (8)$$

to minimize the miss-classification rate.

B. The Pima Indian diabetes data

This data set concerns prediction of Diabetes Mellitus. The included patients are females at least 21 years old of Pima Indian heritage [11]. The data set is split in training and test sets of sizes $N_{\text{train}} = 200$ and $N_{\text{test}} = 332$ respectively. Initial pilot runs indicated that $K = 5$ was a good bias-variance trade-off. The initial training phase on complete data set provided a model with training error rate $E_{\text{train}} = 18.0\%$ and when evaluated on the complete test set showed a test error rate of $E_{\text{test}} = 21.6\%$, at par with test error rates reported elsewhere, see e.g. [12]. To emulate a incomplete data scenario we tested the top-down saliency estimate Eq. (6) using a initial feature vector $\boldsymbol{x}$ comprising variables $(1, 7)$ representing the 'number of times pregnant' and 'age' features within the list of a total of seven measures,

1) number of times pregnant
2) plasma glucose concentration a 2 hours in an oral glucose tolerance test
3) diastolic blood pressure (mm Hg)
4) triceps skin fold thickness (mm)
5) body mass index (weight in kg/(height in m)^2)
6) diabetes pedigree function
7) age (years).

The input feature 2 'Plasma glucose concentration at 2 hours in an oral glucose tolerance test' is interesting for its known diagnostic value [11] and also because it both more expensive to obtain and delayed relative to the other features. The Diabetes Pedigree Function represents the Diabetes Mellitus history in relatives and the genetic relationship of relatives to the subject. It is based on information about parents, grandparents, full and half siblings, full and half aunts and uncles, and first cousins [11], and thus is also complex and time consuming to obtain. To obtain an initial expression of the input feature relevance we measured the mutual information between each input feature and the class label. The mutual information was tested against a null hypothesis of no mutual information using a simple permutation test ($N_{\text{resamples}} = 200$), the null was rejected if $p > 0.01$, and in this case the mutual information was reported. We found that features $(1, 2, 7)$ were significantly informative on a single feature basis, as seen in panel (e) of figure 1, with feature 2 as the most informative as expected. Going through all test examples we note both the saliency allocated to features $(2, 3, 4, 5, 6)$, as well as the rate at which they are chosen for measurement as being the most salient. The resulting distributions are shown in panels (b) and (c) of figure 1. Clearly, feature 2 is the most important in terms of saliency and is attended to in more than 80% of the test data. The choice of features is broken down within the two classes in panel (d), and interestingly we find that global result of panel (b) appears from somewhat different distributions in the two classes indicating the importance of the interaction between the initial input $\boldsymbol{x}$ and the top down mechanism in the saliency estimate.

The classification performances of the various schemes are shown in panel (a). The saliency based feature choice leads to an error rate of 24%, while classification based on the initial features $(1, 7)$ only leads to poor performance close to baseline random guessing. Classification based on $(1, 7)$ combined with a *randomly selected feature* leads to a somewhat higher error rate (29%). In conclusion, we would recommend to acquire the plasma glucose concentration, while the Diabetes Pedigree Function seems irrelevant in the present sample.

C. The Liver Disorders data set

This classification task concerns prediction of liver disorder based on blood tests and alcohol consumption of male individuals. The task is proposed and data donated by BUPA Medical Research Ltd., see [10]. The list of features for this problem comprises

1) mean corpuscular volume,
2) alkaline phosphotase,
3) alamine aminotransferase,
4) aspartate aminotransferase
5) gamma-glutamyl transpeptidase,
6) drinks; as the number of half-pint equivalents of alcoholic beverages drunk per day.

The first 5 variables are blood tests features believed to be sensitive to liver disorders that might arise from alcohol consumption. Each data sample is based on the record of a single male individual [10]. We tested the single feature information content as shown in figure 2 panel (e) and find that only features (5,6) are informative on their own ($p < 0.01$). This data set is smaller with $N_{\text{train}} = 200$ and $N_{\text{test}} = 95$. We train the Gaussian Discrete model with $K = 10$. The baseline error rate on the test data is 42%.

We emulate a severely missing data situation by letting the initial incomplete measurement be given as two features $(3, 4)$ which both are deemed uninformative by the permutation test. As in the diabetes data we run through all test examples and note both the saliency allocated to features $(1, 2, 5, 6)$ as well as the rate at which they are chosen as the most salient. The resulting distributions are shown in panels (b) and (c) of figure 2. The behavioral feature 6 (number of drinks) obtains some attention but is rarely chosen as the additional most salient feature for classification. The most salient and most often chosen features are 1 and 5, that are chosen for attention in 55% and 40% of the test respectively. The choice of features is broken down in classes in panel (d), and again we see that the classes require somewhat different choice of feature showing the interaction between values of the initial input $(3, 4)$ and the top down influence on the saliency estimate.

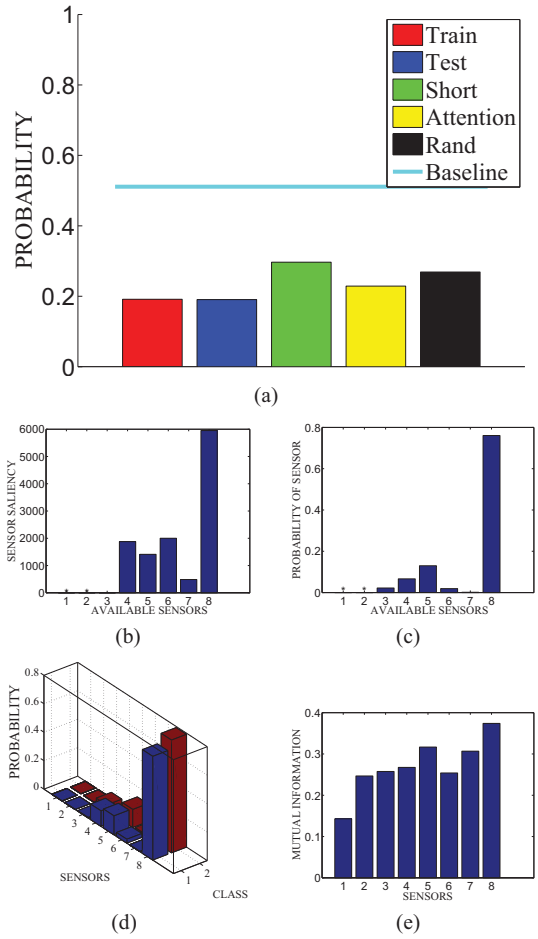


Fig. 3. Abalone data converted to a classification problem (old/young). Eight input features are considered. We simulate incomplete test measurement, in which only features $(1, 2)$ are included. The panels show: (a) Error rates in the training set of complete data ($N_{\text{train}} = 2500$) and the error on the test set ($N_{\text{test}} = 677$) using complete data, test error rate when using the initial feature set $(1, 2)$, test error using $(1, 2)$ and the feature chosen by the top down saliency estimate, and finally the test error opbtained using (1,2) and a randomly chosen additional feature; (b) Estimated information saliency obtained on the test data, given the incomplete feature vector (1,2); (c) Frequency of selection of the additional features; (d) Frequency of selection of features in test cases within the two classes; (e) The $\log_2$ mutual information between features and class label.

We find it interesting that even in the case of such a relatively 'uninformative' initial measurement as we have access to in this experiment, the top down saliency estimate is successful in locating features that lead to a classification performance at par with that obtained when using all features ($\sim 30\%$).

D. The Abalone data

The task is to predict the age of Abalone from physical measurements [13], [14], here we convert the problem to a binary decision problem (young vs. old). The variables used as features are

1) gender (M, F),
2) length, longest shell measurement,
3) diameter, perpendicular to length,
4) height, with meat in shell,
5) whole weight, whole abalone,
6) shucked weight, weight of meat,
7) viscera weight, gut weight (after bleeding),
8) shell weight, after being dried.

This data set is larger with $N_{\text{train}} = 3500$ and $N_{\text{test}} = 677$. We train the Gaussian Discrete model with $K = 17$. The baseline error rate is 50% and the trained model obtains a test error below 20% based on complete measurements, thus this is well calibrated modeling problem. We further tested the single feature information content as shown in figure 3 panel (e) and find that all features are informative ($p < 0.01$), with feature 8 as the most informative with almost 0.4 bits of information. We choose in this case to provide features $(1, 2)$ and evaluate the top down saliency for features $(3, 4, 5, 6, 7, 8)$. As expected, we find feature 8 is allocated most saliency and is most often attended to, being proposed in 75% of the test cases. The resulting classifier is significantly improved over random attention (22% vs. 27%).

E. The Yeast data

This last data set used for illustration of the top down saliency estimate concerns determination of protein cellular localization sites [15]. The variables used as features include

1) McGeoch's method for signal sequence recognition,
2) Von Heijne's method for signal sequence recognition,
3) score of the ALOM membrane spanning region prediction program,
4) score of discriminant analysis of the amino acid content of the N-terminal region (20 residues long) of mitochondrial and non-mitochondrial proteins,
5) presence of "HDEL" substring (thought to act as a signal for retention in the endoplasmic reticulum lumen),
6) peroxisomal targeting signal in the C-terminus,
7) score of discriminant analysis of the amino acid content of vacuolar and extracellular proteins,
8) score of discriminant analysis of nuclear localization signals of nuclear and non-nuclear proteins.

The classification becomes a binary decision process by selecting a subset associated with two most frequent sequence types *CYT* (cytosolic/cytoskeletal 463 examples) and *NUC* (nuclear, 429 examples) in SWISS-PROT database. This data set comprises a training set with $N_{\text{train}} = 650$) and a test set ($N_{\text{test}} = 242$) samples. We train the Gaussian Discrete model with $K = 11$. The baseline error rate is 45% and the trained models obtain test and training error rates around 30% based on complete measurements, thus signifying a more

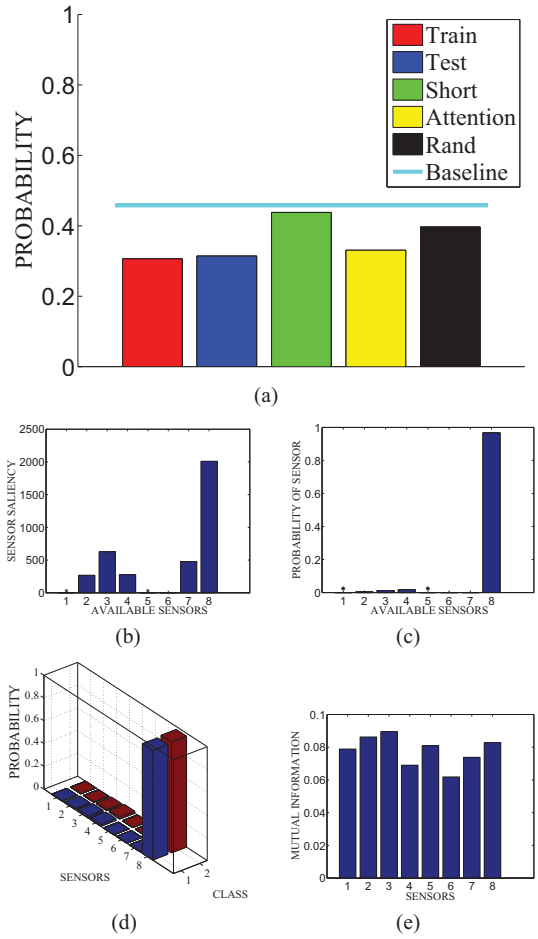


Fig. 4. Yeast data. Eight input features are considered. We simulate incomplete test measurement, in which are given the two features $(1, 5)$. The panels show: (a) Error rates in the training set of complete data ($N_{\text{train}} = 650$) and in the test set ($N_{\text{test}} = 242$) using complete data, test error rate using only the initial feature set $(1, 5)$, and test error using $(1, 5)$, and the feature chosen by the top down saliency estimate, and finally the test error obtained using $(1, 5)$ and a randomly chosen additional feature; (b) Estimated information saliency obtained on the test data, given the incomplete feature vector $(1, 5)$; (c) Frequency of selection of the additional features; (d) Frequency of selection of features in test cases within the two classes; (e) The $\log_2$ mutual information between features and class label.

noisy decision problem than the previous Abalone case. We further tested the single feature information content. As in the Abalone data set all features are informative (see 4 panel (e)). To emulate incomplete data we provide the classifier with features $(1, 5)$ that are both significantly informative, but at a somewhat lower value than in the Abalone example (less than 0.08 bits of information). As in the previous examples we note an improvement in performance of the top down saliency strategy relative to classifying based on the both the initial features $(1, 5)$, as well as compared to providing an extra feature chosen at random. Inspecting the feature selection in figure 4 panel (c) we see that the process almost exclusively chooses to add feature 8 to initially given features $(1, 5)$.

IV. Conclusion

We proposed a computational model of top-down task driven attention. We considered the task to be represented as a classification problem with missing features. Attention is modeled as the choice process over the missing features. The model takes its departure in a classical information theoretic framework for experimental design. This approach requires the evaluation over marginalized and conditional distributions. By implementing the classifier within a Gaussian Discrete mixture it is straightforward to marginalize and condition, hence, we obtained a relatively simple expression for the feature dependent information gain - *the top-down saliency*. As top-down attention is modeled as a simple classification problem, we can evaluate the strategy within conventional classification benchmarks, modified to simulate missing features. We showed in four case studies of well known classification benchmarks from the UCI repository that the attention mechanism provides for improved classification over simple random attention. Improvements were found even if the initial incomplete feature set was of limited information itself.

Acknowledgement

This work is supported in part by the Danish Lundbeck Foundation through the Center for Integrated Molecular Brain Imaging (CIMBI).

References

[1] A. Torralba, M. S. Castelhano, A. Oliva, and J. M. Henderson, "Contextual guidance of eye movements and attention in real-world scenes: the role of global features in object search," *Psychological Review*, vol. 113, pp. 2006, 2006.

[2] S. Ahmad and V. Tresp, "Some solutions to the missing feature problem in vision," in *Advances in Neural Information Processing Systems 5, [NIPS Conference]*, San Francisco, CA, USA, 1993, pp. 393–400, Morgan Kaufmann Publishers Inc.

[3] H. J. Kappen, M. J. Nijman, and M. T. van, "Learning active vision," *Industrial Applications of Neural Networks*, pp. 193–202, 1998.

[4] X. S. Zhou, D. Comaniciu, and A. Krishnan, "Conditional feature sensitivity: A unifying view on active recognition and feature selection," in *ICCV '03: Proceedings of the Ninth IEEE International Conference on Computer Vision*, Washington, DC, USA, 2003, p. 1502, IEEE Computer Society.

[5] W. Wiegerinck, B. Kappen, and W. Burgers, "Bayesian networks for expert systems: Theory and practical applications," in *Interactive Collaborative Information Systems*, pp. 547–578. 2010.

[6] J. van den Berg, A. Curtis, and J. Trampert, "Optimal nonlinear bayesian experimental design: an application to amplitude versus offset experiments," *Geophysical Journal International*, vol. 4, pp. 411–421, 2003.

[7] W. Burgard, D. Fox, and S. Thrun, "Active mobile robot localization by entropy minimization," in *Proc. of the Second Euromicro Workshop on Advanced Mobile Robots*. 1997, IEEE Computer Society Press.

[8] D. V. Lindley, "On a measure of the information provided by an experiment," *Annals Mathematical Statistisc*, vol. 4, pp. 986–1005, 1956.

[9] L. K. Hansen, S. Sigurdsson, T. Kolenda, F. Å. Nielsen, U. Kjems, and J. Larsen, "Modeling text with generalizable gaussian mixtures," in *ICASSP International conference on acoustics, speech and signal processing*, 2000, vol. 4, pp. 3494–3497.

[10] A. Frank and A. Asuncion, "UCI machine learning repository," 2010.
[11] J.W. Smith, J.E. Everhart, W.C. Dickson, W.C. Knowler, and R.S. Johannes, "Using the adap learning algorithm to forecast the onset of diabetes mellitus," in *Proceedings of the Symposium on Computer Applications and Medical Care*. 1988, pp. 261–265, IEEE Computer Society Press.
[12] B. D. Ripley, *Pattern Recognition and Neural Networks*, Cambridge University Press, 1996.
[13] W. J. Nash, T. L. Sellers, S. R. Talbot, A. J. Cawthorn, and W. B. Ford, "The population biology of abalone (haliotis species) in tasmania. i. blacklip abalone (h. rubra) from the north coast and islands of bass strait," 1994, Sea Fisheries Division, Technical Report No. 48 (ISSN 1034-3288).
[14] S. Waugh, "Extending and benchmarking cascade-correlation," 1995, PhD thesis, Computer Science Department, University of Tasmania.
[15] P. Horton and K. Nakai, "A probablistic classification system for predicting the cellular localization sites of proteins," 1996, pp. 109–115, St. Louis, USA.

Appendix D

Top-down attention with features missing at random

S.G. Karadogan , L. Marchegiani, J. Larsen and L.K. Hansen. Top-down attention with features missing at random. *IEEE International Workshop on Machine Learning For Signal Processing*, pp. 1-6, 2011.

TOP-DOWN ATTENTION WITH FEATURES MISSING AT RANDOM

Seliz G. Karadoğan[1] *Letizia Marchegiani*[1,2] *Jan Larsen*[1] *Lars Kai Hansen*[1,3]

[1] DTU Informatics, Technical University of Denmark,
DK-2800, Kgs. Lyngby, Denmark
[2] Department of Computer and System Sciences,
Sapienza, University of Rome, 00185 Roma, Italy
[3] Department of Signal Theory and Communications,
Universidad Carlos III de Madrid, 28911 Leganes, Spain

ABSTRACT

In this paper we present a top-down attention model designed for an environment in which features are missing completely at random. Following (Hansen et al., 2011) we model top-down attention as a sequential decision making process driven by a task - modeled as a classification problem - in an environment with random subsets of features missing, but where we have the possibility to gather additional features among the ones that are missing. Thus, the top-down attention problem is reduced to finding the answer to the question *what to measure next?* Attention is based on the top-down saliency of the missing features given as the estimated difference in classification confusion (entropy) with and without the given feature. The difference in confusion is computed conditioned on the available set of features. In this work, we make our attention model more realistic by also allowing the initial training phase to take place with incomplete data. Thus, we expand the model to include a missing data technique in the learning process. The top-down attention mechanism is implemented in a Gaussian Discrete mixture model setting where marginals and conditionals are relatively easy to compute. To illustrate the viability of expanded model, we train the mixture model with two different datasets, a synthetic data set and the well-known Yeast dataset of the UCI database. We evaluate the new algorithm in environments characterized by different amounts of incompleteness and compare the performance with a system that decides next feature to be measured at random. The proposed top-down mechanism clearly outperforms random choice of the next feature.

Index Terms— Machine learning, missing data techniques, attention modeling, entropy

1 Introduction

In everyday life the human brain is continuously hit by a mixture of stimuli coming from different directions and of a different nature. Most of these signals are irrelevant to the operation of the brain and can be ignored. Attention mechanisms suggest which stimuli have to be processed and which ones can be ignored. This selection is influenced by several factors in an environment and on the eventual presence of top level goals. In particular, attention can be driven by salient signals 'popping up' of the environment, because of the characteristics of the brain's receptive fields (*bottom-up attention*) or by the intentions of the subject to accomplish a task (*top-down attention*).

We are interested in modeling the latter, i.e., knowledge based top-down driven selection procedures. Our model departs from a scenario in which an initial incomplete measurement is available and we want to know what the next measurement is to perform, among those possible, for obtaining as much information as we can to execute the task (*sequential measurement problem*). The aim is to take advantage of optimization strategies to increase our knowledge and, consequently, improve the decision making process. Meanwhile, we highlight the fact that our attention model that is based on a top-down driven feature selection mechanism among missing ones should not be confused with feature selection methods which aim to reduce number of redundant features especially for large databases. Related aims have been pursued in a number of research fields like robot navigation [1], active vision [2, 3], petro-physics and medical diagnosis [4] and geophysics [5]. Hitherto, the solutions proposed have been quite strongly bound to the context, however, in [6] we proposed a formulation which implements top-down task driven attention as a generic classification decision problem allowing us to make more general inferences about the problem. In this paper, our aim is to elaborate further on this approach and expand it to cope with features missing at random.

The decision process we expose is relative to a classification problem, based on a Gaussian Mixture model previously trained on the incomplete measurements available. In [6] we trained a generative classification model, exploiting a complete training set and then, we use it for testing data sets with missing features. In particular, we considered a deterministic setup in which a given set of features were missing in all cases in the test set. In this work, instead, both the training and test sets have missing features and the incompleteness in both cases (training and test sets) is random. Specifically, according to the definition given by Rubin in [7], the data are *missing completely at random* (MCAR), since the probability of incompleteness depends on neither the observed nor the missing data.

Different missing data techniques have been analyzed and proposed in different areas to handle the missing data problem. According to the way in which these techniques deal with incomplete information, it is possible to divide them in three major categories: *deletion* methods, *imputation* methods and *model-based* methods. In the first category, only the present data are used. In the imputation methods, all the accessible information is held and utilized and the 'holes' in the data set are substituted by other estimates. In the *multiple imputation (MI)*, proposed by Rubin in [7], several values are inserted for each missing datum. The model-based methods like, e.g., maximum Likelihood approaches [8], we do not modify or ignore part of the available data, but operate directly on the incomplete set.

In [9] we evaluated a number of missing data techniques on classification problems, computing the misclassification rate (MR) for different missing data percentages (MDP) and proposing a new way

to quantify their efficiency. In particular, we analyzed the behavior of *mean imputation* method, in which the missing variable is replaced with the mean value of the same variable, the *pairwise deletion* method, which consists of removing only the missing elements, and the maximum likelihood method in [8] (*Complete EM*). The results showed that Complete EM is the most efficient both in cases of complete and with missing data test sets leading us to apply this methods also in this work.

We use the method proposed in [8] to train and test the Gaussian discrete mixture model. In particular, we compute the saliency of the missing data conditional on the available ones. The saliency is given by the difference in confusion (*entropy*) before and after the measurement of the additional chosen features. We evaluate on classification problems and compare our saliency based method of selecting features with random selection. The results obtained show that, even considering a more complex scenario due to the data missing completely at random and the possibility to have an incomplete training set to estimate the model, the misclassification rate is still reduced when we base the selection of the additional feature on estimated information gain compared to random selection.

The paper is organized as follows, in the next section we explain the machine learning model, the missing data technique used and the attention model. In Section 3 we explain the experimental setup and give the results of two different classification problems in which data is MCAR and we apply the attention model as a solution to sequential measurement problem. Finally we provide concluding remarks in Section 4.

2 Modeling Framework

2.1 Modeling Framework

We implement the top-down attention mechanism within the Gaussian Discrete mixture model, see e.g., [10]. Define $\boldsymbol{x}$ as the d-dimensional input feature vector and the associated output, $y \in \{1, 2, \cdots, C\}$, of class labels, assuming C mutually exclusive classes. The joint input/output density is modeled as

$$p(y, \boldsymbol{x}|\boldsymbol{\theta}) = \sum_{k=1}^{K} P(y|k)p(\boldsymbol{x}|k)P(k) \tag{1}$$

$$p(\boldsymbol{x}|k) = \frac{1}{\sqrt{|2\pi\boldsymbol{\Sigma}_k|}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_k)^\top \boldsymbol{\Sigma}_k^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_k)\right), \tag{2}$$

where K is the number of components, $p(\boldsymbol{x}|k)$ are the component Gaussians mixed with the non-negative proportions $P(k)$, $\sum_{k=1}^{K} P(k) = 1$ and the class-cluster conditionals re denoted by $P(y|k)$. $\boldsymbol{\theta}$ is the vector of all model parameters, i.e., $\boldsymbol{\theta} \equiv \{P(y|k), \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, P(k) : \forall k, y\}$. The joint input/output for each components is assumed to factorize, i.e., $p(y, \boldsymbol{x}|k) = P(y|k)p(\boldsymbol{x}|k)$. The input density associated with Eq. (1) is given by

$$p(\boldsymbol{x}|\boldsymbol{\theta}_u) = \sum_{y=1}^{C} p(y, \boldsymbol{x}) = \sum_{k=1}^{K} p(\boldsymbol{x}|k)P(k),$$

where $\boldsymbol{\theta}_u \equiv \{\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, P(k) : \forall k, y\}$. Assuming a 0/1 loss function the optimal Bayes' classification rule is $\widehat{y} = \max_y P(y|\boldsymbol{x})$ where[1]

$$P(y|\boldsymbol{x}) = \frac{p(y, \boldsymbol{x})}{p(\boldsymbol{x})} = \sum_{k=1}^{K} P(y|k)P(k|\boldsymbol{x}),$$

[1]The dependence on $\boldsymbol{\theta}$ is suppressed for clarity.

with $P(k|\boldsymbol{x}) = p(\boldsymbol{x}|k)P(k)/p(\boldsymbol{x})$.

Assume that we are given a data set of independent i.i.d. input-output examples $\mathcal{D} = \{\boldsymbol{x}_n, y_n; n = 1, 2, \cdots, N\}$. The negative log-likelihood is given by

$$L = -\log p(\mathcal{D}|\boldsymbol{\theta}) = -\sum_{n\in\mathcal{D}} \log \sum_{k=1}^{K} P(y_n|k)p(\boldsymbol{x}_n|k)P(k)$$

We use an iterative modified EM algorithm to estimate the model parameters [10]:

1. For initialization we model the data as a normal distribution with parameter $\boldsymbol{\mu}_0$ and $\boldsymbol{\Sigma}_0$. Only observed data is used during calculations in the case of missing data. However, $\boldsymbol{\Sigma}_0$ is regularized since the estimated covariance matrix is unbiased and is not guaranteed to be positive semi definite. The regularization is based on inflating the diagonal elements similar to the approach presented in [11]. Further details can be found in [9]. We draw random samples from the distribution, $\boldsymbol{\mu}_0$ and $\boldsymbol{\Sigma}_0$ and apply the KKZ method [12]. Last, $P(k) = 1/K$ and $P(y|k)$ is randomly initialized.
2. Compute posterior component probability for all $n \in \mathcal{D}$:
$$P(k|y_n, \boldsymbol{x}_n) = \frac{P(y_n|k)p(\boldsymbol{x}_n|k)P(k)}{\sum_k P(y_n|k)p(\boldsymbol{x}_n|k)P(k)}. \tag{3}$$
3. For all k, update mean vectors and covariance matrices
$$\boldsymbol{\mu}_k = \frac{\sum_{n\in\mathcal{D}} \boldsymbol{x}_n P(k|y_n, \boldsymbol{x}_n)}{\sum_{n\in\mathcal{D}} P(k|y_n, \boldsymbol{x}_n)}, \quad \boldsymbol{\Sigma}_k = \frac{\sum_{n\in\mathcal{D}} \boldsymbol{S}_{kn} P(k|y_n, \boldsymbol{x}_n)}{\sum_{n\in\mathcal{D}} P(k|y_n, \boldsymbol{x}_n)}$$
where $\boldsymbol{S}_{kn} = (\boldsymbol{x}_n - \boldsymbol{\mu}_k)(\boldsymbol{x}_n - \boldsymbol{\mu}_k)^\top$.
4. For all k update cluster priors and class cluster posteriors
$$P(k) = \frac{\sum_{n\in\mathcal{D}} P(k|y_n, \boldsymbol{x}_n)}{N}, \quad P(y|k) = \frac{\sum_{n\in\mathcal{D}} \delta_{y-y_n} P(k|y_n, \boldsymbol{x}_n)}{\sum_{n\in\mathcal{D}} P(k|y_n, \boldsymbol{x}_n)}$$

2.2 Missing Data Technique:Complete Expectation Maximization

The missing data technique that we apply for both training and test set is the Complete Expectation Maximization (CEM) [9] since in our earlier work we showed that it is an efficient method compared to mean imputation, pairwise-listwise deletion methods [9]. This method is a maximum likelihood missing data technique originally proposed in [8]. Missing data compensation is carried out within the EM step of the model where also the model components are estimated. The posterior component probability, $p(k|y_n, \boldsymbol{x}_n)$, is again calculated as in Eq. (3), but only on observed dimensions. For updating the mean vector, $E[\boldsymbol{x}_n^m|\boldsymbol{x}_n^o]$ ($\boldsymbol{x}_n^m$ and $\boldsymbol{x}_n^o$ represent missing and observed parts of $\boldsymbol{x}_n$) substitutes missing components of $\boldsymbol{x}_n$, and for updating the covariance matrix, $E[\boldsymbol{x}_n^m \boldsymbol{x}_n^{m^T}|\boldsymbol{x}_n^o]$ is substituted for outer product matrices containing missing components

$$E[\boldsymbol{x}_n^m|\boldsymbol{x}_n^o] = \mu_k^m + \boldsymbol{\Sigma}_k^{mo}{\boldsymbol{\Sigma}_k^{oo}}^{-1}(\boldsymbol{x}_n^o - \mu_k^o),$$

$$E[\boldsymbol{x}_n^m \boldsymbol{x}_n^{m^\top}|\boldsymbol{x}_n^o] = \boldsymbol{\Sigma}_k^{mm} - \boldsymbol{\Sigma}_k^{mo}{\boldsymbol{\Sigma}_k^{oo}}^{-1}\boldsymbol{\Sigma}_k^{mo^T} + E[\boldsymbol{x}_n^m|\boldsymbol{x}_n^o]E[\boldsymbol{x}_n^m|\boldsymbol{x}_n^o]^\top.$$

2.3 Attention Model

We apply the top-down attention mechanism that was recently proposed in [6] and for completeness we provide the derivation of the input saliency below.

There are many studies on missing data techniques that enable us to train and classify with missing input features, see e.g., [13]. However, the sequential measurement problem, i.e., which measurement to make next with incomplete data, is not studied much. Kappen et al. [2] considered the sequential measurement process as a two-stage process as we do in this work. As above, we represent the signal detection problem by a probability distribution over a set of C classes indexed by the discrete variable y, $(y = 1, ..., C)$. However, we split the input data in observed and un-observed as follows. Initially the observation $\boldsymbol{x}$ with components $x_i,\ i = 1, ..., I$ is available. Additional measurement z_j is chosen among the set of missing features $\boldsymbol{z}$ with components $z_1, ..., z_J$ in the second step of the process.

Denote the joint probability of the classes and all features observed and missing as $p(y, \boldsymbol{x}, \boldsymbol{z})$. We use Bayesian decision theory and use the posterior distribution $p(y|...)$ to make inferences about y. The condition depends on the stage in the sequential measurement process. Initially, the information available is $\boldsymbol{x}$, thus the relevant probability is

$$\begin{aligned} p(y|\boldsymbol{x}) &= \int p(y, \boldsymbol{z}|\boldsymbol{x}) d\boldsymbol{z} \\ &= \frac{\int p(y, \boldsymbol{x}, \boldsymbol{z}) d\boldsymbol{z}}{\sum_{y=1}^{C} \int p(y, \boldsymbol{x}, \boldsymbol{z}) d\boldsymbol{z}}. \end{aligned} \tag{4}$$

We select an additional feature $\boldsymbol{z}_j$ using the top down attention model, that will result in the distribution

$$\begin{aligned} p(y|\boldsymbol{x}, z_j) &= \sum_{y=1}^{C} \int p(y, \boldsymbol{z}|\boldsymbol{x}) \prod_{i \neq j} dz_i \\ &= \frac{\int p(y, \boldsymbol{x}, \boldsymbol{z}) \prod_{i \neq j} dz_i}{\sum_{y=1}^{C} \int p(y, \boldsymbol{x}, \boldsymbol{z}) \prod_{i \neq j} dz_i} \end{aligned} \tag{5}$$

The difference in confusion (entropy) before and after the second measurement is the information obtained by the additional measurement, and this is the quantity we want to maximize

$$\begin{aligned} \Delta S_j(\boldsymbol{x}, z_j) &= \sum_{y=1}^{C} \log p(y|\boldsymbol{x}, z_j) p(y|\boldsymbol{x}, z_j) \\ &- \sum_{y=1}^{C} \int \log p(y, \boldsymbol{z}|\boldsymbol{x}) p(y, \boldsymbol{z}|\boldsymbol{x}) d\boldsymbol{z} \end{aligned} \tag{6}$$

Since we do not know z_j at this step, we average $\Delta S_j(\boldsymbol{x}, z_j)$ with respect to this variable given the information we have, i.e., with respect to the distribution of z_j conditioned on the initial measurement $\boldsymbol{x}$. This procedure provides us with the *expected information gain* of measuring the value of feature j,

$$\begin{aligned} G_j(\boldsymbol{x}) &\equiv \int \Delta S_j(\boldsymbol{x}, z_j) p(z_j|\boldsymbol{x}) dz_j \\ &= \sum_{y=1}^{C} \int \log p(y|\boldsymbol{x}, z_j) p(y, z_j|\boldsymbol{x}) dz_j \\ &\quad - \sum_{y=1}^{C} \int \log p(y, \boldsymbol{z}|\boldsymbol{x}) p(y, \boldsymbol{z}|\boldsymbol{x}) d\boldsymbol{z}. \end{aligned} \tag{7}$$

The information gain can be used to rank features in importance [2, 14], hence the name 'top-down saliency'. The second term can be neglected in the saliency estimate since it does not depend on j.

Introducing the Gaussian discrete model (Eq. (1)) within the expression for the information gain and using $p(k|\boldsymbol{x}) = p(k)p(\boldsymbol{x}|k)/p(\boldsymbol{x})$, we obtain

$$\begin{aligned} G_j(\boldsymbol{x}) &= \sum_{y=1}^{C} \sum_{k=1}^{K} p(y|k) p(k|\boldsymbol{x}) \times \\ &\qquad \int \log \left[p(y, \boldsymbol{x}, z_j) \right] p(z_j|\boldsymbol{x}, k) dz_j \\ &- \sum_{k=1}^{K} p(k|\boldsymbol{x}) \int \log \left[p(\boldsymbol{x}, z_j) \right] p(z_j|\boldsymbol{x}, k) dz_j \\ &+ \text{const.} \end{aligned} \tag{8}$$

where $p(y, \boldsymbol{x}, z_j) = \sum_{k=1}^{K} p(k) p(y|k) p(\boldsymbol{x}, z_j|k)$ and $p(\boldsymbol{x}, z_j) = \sum_{k=1}^{K} p(k) p(\boldsymbol{x}, z_j|k)$. Thus, computing G_j for all I features amounts to computing $Q = I * (C + 1) * K$ one-dimensional integrals over Gaussian measures $p(z_j|\boldsymbol{x}, k) = \mathcal{N}(\mu_j(\boldsymbol{x}, k), \sigma_j^2(\boldsymbol{x}, k))$ with

$$\begin{aligned} \mu_j(\boldsymbol{x}, k) &= \mu_{j,k} + \boldsymbol{\Sigma}_{z_j, \boldsymbol{x}, k} \boldsymbol{\Sigma}_{\boldsymbol{x}, \boldsymbol{x}, k}^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_{\boldsymbol{x}, k}) \\ \sigma_j^2(\boldsymbol{x}, k) &= \sigma_{j,k}^2 - \boldsymbol{\Sigma}_{z_j, \boldsymbol{x}, k} \boldsymbol{\Sigma}_{\boldsymbol{x}, \boldsymbol{x}, k}^{-1} \boldsymbol{\Sigma}_{\boldsymbol{x}, z_j, k}. \end{aligned} \tag{9}$$

In these expressions $\mu_{j,k}, \sigma_{j,k}^2$ are the mean and variance of the jth feature in the kth component, while $\boldsymbol{\Sigma}_{a,b,k}$ is the part of the covariance matrix of the k component corresponding to variable sets a, b.

3 Experimental Evaluations

The main aim of our algorithm is to reduce the error rate of an ensuing decision classification problem. Hence, we evaluate the model by the misclassification rates on test data. To illustrate the approach, we generate synthetic data with known distributional properties that conform with the model, and we analyze a well-known benchmark problem from the UCI depository [15], often referred to as the 'Yeast dataset'.

We first simulate an incomplete measurement situation on N_train and N_test data points. Data is missing completely at random with different missing data percentages. The randomization is done such that not all values can be missing in one observation (One feature per observation is kept present randomly). We test the performance either choosing the next feature to be measured with highest saliency or randomly and compare with classifiers, one trained with full data and the other the with original missing data we have.

The experiments are carried out using MATLAB. The Gaussian mixture model is trained and initialized as explained in Section 2 with a multi-start procedure with 5 different initializations, and 250 iterations.

3.1 Synthetic Dataset

Synthetic data is generated by a Gaussian mixture model to test the algorithm. The number of mixtures K, is 3, and the number of dimensions D is 10. The difficulty of the problem is determined with an SNR calculation based on a distance measure defined between clusters of data.

Let ds_{kl} be the distance between μ_k and μ_l, eig_k be a vector consisting of eigenvalues of Σ_k and mean() be the arithmetic mean operator, then

$$\text{SNR}_{\text{dB}} = 10 \log \left(\frac{\sum_{1 \leq \text{k} \leq \text{K}, \text{k} \leq \text{l} \leq \text{K}} \text{ds}_{\text{kl}}^2}{\sum_{1 \leq \text{k} \leq \text{K}} \text{mean}(\text{eig}_\text{k})} \right) \tag{10}$$

Figure 1 shows the illustration of our SNR calculation for a 2D data with 3 clusters. If we formulate that SNR calculation example according to the equation above,then we have:

$$\mathrm{SNR}_{\mathrm{dB}} = 10\log\left(\frac{\mathrm{ds}_{12}^2 + \mathrm{ds}_{13}^2 + \mathrm{ds}_{23}^2}{(\mathrm{a}_1^2 + \mathrm{b}_1^2)/2 + (\mathrm{a}_2^2 + \mathrm{b}_2^2)/2 + (\mathrm{a}_3^2 + \mathrm{b}_3^2)/2}\right)$$

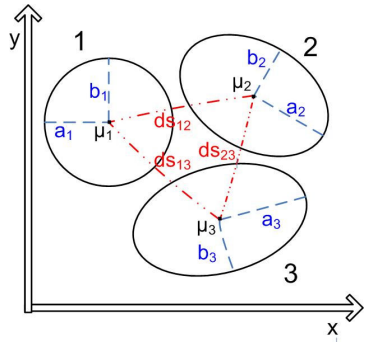


Fig. 1: The illustration of SNR calculation for a 2D data with 3 clusters

$N\_\mathrm{train} = 1500$ and $N\_\mathrm{test} = 300$ observations are generated for training and test sets respectively. We check the performance in the interval from 40% to 70%. The baseline error rate is 66% and the trained models with missing data have error rates between 2% and 15% based on complete test measurements.

We use an SNR of 10 dB (check the equation 10) for the main experiments, though we experiment also with data with different SNRs to observe the effect of problem complexity. Figure 2 shows the first three principal components plotted against each other for this dataset, for 15 dB and 5 dB of SNRs as an example.

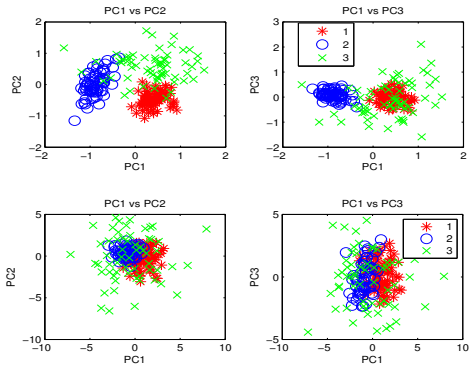


Fig. 2: The principal components (PCs) plot for the data generated, with SNR of 15 dB (top) and 5 dB (bottom).

We designed the data set such that all features are equally important. We measure the mutual information between each input feature and the class label. We use a permutation test ($N_{\mathrm{resamples}} = 200$) to test the mutual information against a null hypothesis of no mutual information. Mutual information is recorded if the null is rejected with $p > 0.01$. We find that all features are informative as expected (see Figure 3 (b)). Figure 3 (c) shows the frequency of which features are selected with the attention model. We observe that the frequencies are not simply given by the mutual information. This result underlines the need for an attention model. In addition, we observe that the feature saliency depends on the class label as well (see Figure 3 (d)). Figure 3 (a) shows the error rates for the different methods for variable SNR. For all situations, we train the model with missing data (MCAR). We test with missing data and full data where all features are available. We compare the performances of those to the cases in which we test with missing data with one additional feature chosen randomly or using the attention model. As expected, we reduce the error rate adding one feature randomly or with attention model. However, the top-down attention model outperforms the random saliency model for all MDPs being closer to the full data error rate.

We investigate how well the attention model performs for problems with different complexities. We change the difficulty of the problem by generating data with different SNRs (see equation 10, the higher the SNR the easier the problem, see also Figure 2). We observe that the attention model is less effective for the more difficult problems (Figure 4).

3.2 The Yeast Dataset

The Yeast data set used for the analysis of our top-down attention model concerns determination of protein cellular localization sites [16].

The features include

1. McGeoch's method for signal sequence recognition,
2. Von Heijne's method for signal sequence recognition,
3. score of the ALOM membrane spanning region prediction program,
4. score of discriminant analysis of the amino acid content of the N-terminal region (20 residues long) of mitochondrial and non-mitochondrial proteins,
5. presence of "HDEL" substring (thought to act as a signal for retention in the endoplasmic reticulum lumen),
6. peroxisomal targeting signal in the C-terminus,
7. score of discriminant analysis of the amino acid content of vacuolar and extracellular proteins,
8. score of discriminant analysis of nuclear localization signals of nuclear and non-nuclear proteins.

We select a subset associated with two most frequent sequence types *CYT* (cytosolic/cytoskeletal 463 examples) and *NUC* (nuclear, 429 examples) in SWISS-PROT database reducing the classification problem to a binary decision. We have a training set with ($N_{\mathrm{train}} = 630$) and a test set ($N_{\mathrm{test}} = 262$) samples. We train the Gaussian Discrete model with $K = 11$. The baseline error rate is 50% and the trained models with missing data give error rates around 40% based on complete test measurements. This is a more noisy decision problem than the synthetic one.

Unlike the synthetic data, not all features are informative, while features (1,8) are the most informative ones (see Figure 5 (b)). Even if there are similarities between the mutual information and the selection of features with attention model (see Figure 5 (b) and (c)) (feature 8 is the most informative and the most frequently selected), we observe the features that are not informative are selected quite

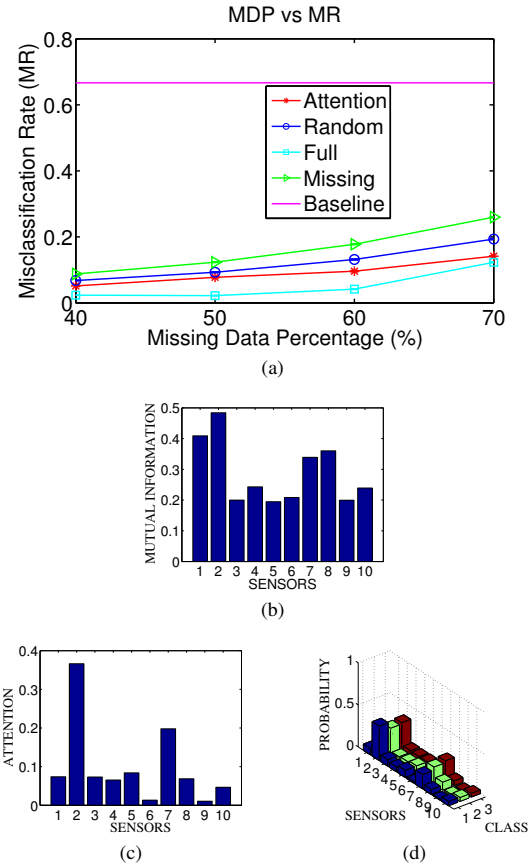


Fig. 3: Synthetic data. Ten input features are considered. Train and test data are missing completely at random (MCAR). (a) For different missing data percentages, the misclassification error rates for the test set where data is MCAR (Missing), all missing features are added (Full), one random feature is added among missing (Random) and one feature is added among missing using the attention model (Attention). (b) The log_2 mutual information between features and class label. (c) Frequency of selection of additional features with attention model. (d) Frequency of selection of features within the three classes.

often as well. This result again underlines the need for an attention model that combine the available data and the task.

The saliency of the features depends on the class label like in synthetic data (see Figure 5 (d)). Figure 5 (a) shows the error rates for MDP of 50% (not all MDPs were convenient to use, too few or too much data problem) for the same situations as explained for synthetic data (data is MCAR, full, missing with added feature randomly or with attention model). The attention model performs better than choosing a random feature to be measured.

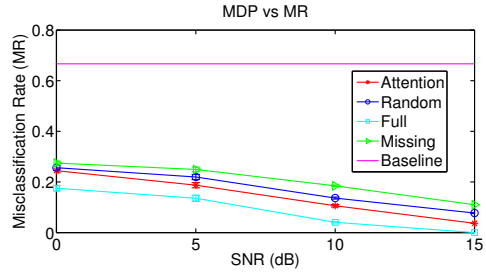


Fig. 4: The result for syhthetic dataset created with different SNRs. The easier the problem (higher SNR), the stronger the effect of our attention modeling

4 Conclusion

We proposed a top-down attention model with features missing completely at random. Attention was modeled as sequential decision making over the missing features. The saliency of features was given as the difference in confusion (entropy). By using a Gaussian mixture as the learning model we obtained a relatively simple expression for the feature dependent information gain. Both train and test data were assumed to be incomplete within this work. To compensate for the missing data, we applied a missing data technique that was already proven to be an efficient method [9]. We applied the method to two classification problems (synthetic and Yeast datasets) and we showed that the top-down attention model we proposed outperformed simple random attention. We investigated the dependency of the method to the complexity of the problem (problems with different SNRs and different missing data percentages) and showed that the potential of the method is greatest for easy problems with higher missing data rate (the complexity of the problem should be analyzed to decide for the convenience of the use of the method).

Acknowledgement

This work is supported in part by the Danish Lundbeck Foundation through the Center for Integrated Molecular Brain Imaging (CIMBI).

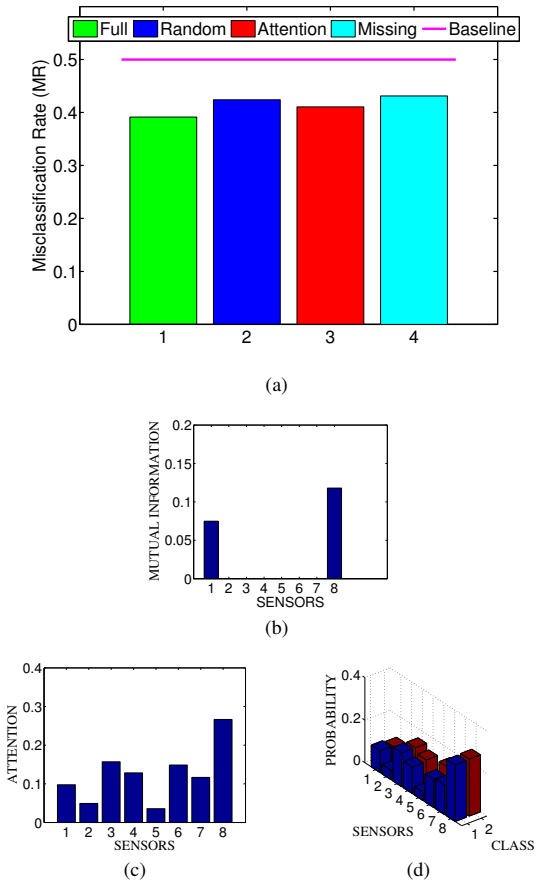


Fig. 5: Yeast data. Eight input features are considered. Train and test data are missing completely at random (MCAR). (a) For 50% of missing data percentage , the misclassification error rates for the test set where data is MCAR (Missing), all missing features are added (Full), one random feature is added among missing (Random) and one feature is added among missing using the attention model (Attention). (b) The log_2 mutual information between features and class label. (c) Frequency of selection of additional features with attention model. (d) Frequency of selection of features within the two classes.

5 References

[1] W. Burgard, D. Fox, and S. Thrun, "Active mobile robot localization by entropy minimization," in *Proc. of the Second Euromicro Workshop on Advanced Mobile Robots*. 1997, IEEE Computer Society Press.

[2] H. J. Kappen, M. J. Nijman, and M. T. van, "Learning active vision," *Industrial Applications of Neural Networks*, pp. 193–202, 1998.

[3] X. S. Zhou, D. Comaniciu, and A. Krishnan, "Conditional feature sensitivity: A unifying view on active recognition and feature selection," in *ICCV '03: Proceedings of the Ninth IEEE International Conference on Computer Vision*, Washington, DC, USA, 2003, p. 1502, IEEE Computer Society.

[4] W. Wiegerinck, B. Kappen, and W. Burgers, "Bayesian networks for expert systems: Theory and practical applications," in *Interactive Collaborative Information Systems*, pp. 547–578. 2010.

[5] J. van den Berg, A. Curtis, and J. Trampert, "Optimal nonlinear bayesian experimental design: an application to amplitude versus offset experiments," *Geophysical Journal International*, vol. 4, pp. 411–421, 2003.

[6] L.K. Hansen, S.G. Karadogan, and L. Marchegiani, "What to measure next to improve decision making? On top-down task driven feature saliency," in *SSCI, IEEE Symposium Series on Computational Intelligence*, 2011, Accepted.

[7] D.B. Rubin, *Multiple Imputations for nonresponse in surveys*, New York: Wiley, 1987.

[8] Z. Ghahramani and M.I. Jordan, "Supervised learning from incomplete data via an EM approach," in *Advances in Neural Information Processing Systems 6*. 1994, pp. 120–127, Morgan Kaufmann.

[9] S.G. Karadogan, L. Marchegiani, L.K. Hansen, and J. Larsen, "How efficient is estimation with missing data?," in *ICASSP, International Conference on Acoustics, Speech and Signal Processing*, 2011, Accepted.

[10] L.K. Hansen, S. Sigurdsson, T. Kolenda, F.A. Nielsen, U. Kjems, and J. Larsen, "Modeling text with generalizable Gaussian mixtures," in *ICASSP'00. Proceedings*. IEEE, 2000, vol. 6, pp. 3494–3497.

[11] T. Schneider, "Analysis of incomplete climate data: Estimation of mean values and covariance matrices and imputation of missing values," *Journal of Climate*, vol. 14, no. 5, pp. 853–871, 2001.

[12] T. Su and J.G. Dy, "In search of deterministic methods for initializing K-means and Gaussian mixture clustering," *Intelligent Data Analysis*, vol. 11, no. 4, pp. 319–338, 2007.

[13] S. Ahmad and V. Tresp, "Some solutions to the missing feature problem in vision," in *Advances in Neural Information Processing Systems 5, [NIPS Conference]*, San Francisco, CA, USA, 1993, pp. 393–400, Morgan Kaufmann Publishers Inc.

[14] D. V. Lindley, "On a measure of the information provided by an experiment," *Annals Mathematical Statistisc*, vol. 4, pp. 986–1005, 1956.

[15] A. Frank and A. Asuncion, "UCI machine learning repository," 2010.

[16] P. Horton and K. Nakai, "A probablistic classification system for predicting the cellular localization sites of proteins," 1996, pp. 109–115, St. Louis, USA.

Appendix E

The Role of Top-Down Attention in the Cocktail Party: Revisiting Cherry's Experiment after Sixty Years

---

L.Marchegiani , S.G. Karadogan , T. Andersen, J.Larsen and L. K. Hansen. The Role of Top-Down Attention in the Cocktail Party: Revisiting Cherry's Experiment after Sixty Years. *International Conference on Machine Learning and Applications (ICMLA)*, pp. 183-188, 2011.

The Role of Top-Down Attention in the Cocktail Party: Revisiting Cherry's Experiment after Sixty Years

Letizia Marchegiani*†, Seliz G. Karadoğan*, Tobias Andersen*, Jan Larsen* and Lars Kai Hansen*‡
*DTU Informatics, Technical University of Denmark, DK-2800, Kgs. Lyngby, Denmark
Email: malet,seka,ta,jl,lkh@imm.dtu.dk
†Department of Computer and System Sciences,
Sapienza, University of Rome, 00185 Roma, Italy
‡Department of Signals and Communications,
University Carlos III, Madrid, Spain

***Abstract*—We investigate the role of top-down task drive attention in the cocktail party problem. In a recently proposed computational model of top-down attention it is possible to simulate the cocktail party problem and make predictions about sensitivity to confounders under different levels of attention. Based on such simulations we expect that under strong top-down attention pattern recognition is improved as the model can compensate for noise and confounders. We next investigate the role of temporal and spectral overlaps and speech intelligibility in humans, and how the presence of a task influences their relation. For this purpose, we perform behavioral experiments inspired by Cherry's classic experiments carried out almost sixty years ago. We make participants listen to a mono signal consisting of two different narratives pronounced by a speech synthesizer under two different conditions. In the first case, participants listen with no specific task, while in the second one they are asked to follow one of the stories. Participants report the words they heard by choosing from a list which also includes terms not present in any of the narratives. We define temporal and spectral overlaps using the ideal binary mask (IBMs) as a gauge. We analyze the correlation between overlaps and the amount of reported words. We observe a significant negative correlation when there is no task, while no correlation is detected when a task is involved. Hence, results that are well aligned with the simulation results in our computational top-down attention model.**

I. Introduction

The cocktail party is an often used analogy in machine learning and signal processing, referring to the situation in which multiple signals are mixed and the aim is to separate these, or to recover at least one of the signals in the mixture. In the case of audio mixtures humans are very efficient cocktail party solvers, using a multitude of cues including spatial, spectral, and content as was demonstrated in the famous experiments carried out by Colin Cherry almost sixty years ago [1].

We note that many machine learning applications face 'cocktail parties'. In biomedical signals noise and confounders are often structured and share features with the wanted signal, hence, prior information about the signals and the ways they are mixed plays an important role [2]. Modern telecommunication systems are critically dependent on the ability to recover individual signals from spread spectrum mixtures [3]. Most of the general methods use low level statistical properties of the signals, such as independence [4], or other simple distributional assumptions.

We are interested in the audio cocktail party problem, in particular, the 'hard version' considered in [1], in which, conventional audio cues (spatial and spectral) are removed and the solution is based on high level features related to content. As discussed in [1], this problem can help shed light on the mechanisms applied by human cognition. We are particularly interested in the influence of top-down attention [5], [6].

The audio cocktail party problem is also of great practical importance, e.g., in the transcription of multi-speaker conferences, meetings, seminars and in dialogue systems of robots operating in complex acoustic scenes. Automatic speech recognition procedures need to isolate the voice of interest within confounding sounds and voices around and to track it, to be able to recognize the words pronounced ([7], [8]).

Experiments have been conducted to investigate which characteristics of the auditory scene could help the segregation process of a mixture of stimuli and in which way they can influence each other and the human ability to discriminate within the different signals. The majority of these experiments make use of pure tones, see e.g. [9]). But there are also cases in which human auditory behaviours are tested in situations with multiple speakers, see e.g., [9], [10] and [11]. Sound fission seems to be more pronounced with the voices of the same gender [11], with sounds emitted from close positions in the scene or sounds having enough difference in their fundamental frequencies, in their phase spectrum or in their intensities [10]. Meanwhile, also the vocal tract size, accent or other prosodic features can change the complexity of the grouping of signals belonging to the same stream [9]. For a more complete review, see [12] and the more recent [13].

Bregman and others ([9], [14], [15]), argue that the way in which stimuli are perceived as part of the same flow (coming from the same acoustic source), and the proficiency in selecting just one of these flows and understanding its content, is widely affected by attention, both on a bottom-up and a top-down perspective. It should also be considered that, in fact, the segregation ability is a learned skill and is improved by experience. In psychoacoustics masking refers to the effect that one signal prohibits the other from being detected. Brungart et al. make a distinction between energetic and informational masking: *"Traditional energetic masking occurs when both utterances contain energy in the same critical bands at the same time and portions of one or both of the speech signals are rendered inaudible at the periphery. Higher-level informational masking occurs when the signal and masker are both audible but the listener is unable to disentangle the elements of the target signal from a similar-sounding distracter"* [16].

In order to better understand top-down attention, and how it may modulate informational masking effects we return to Cherry's basic experimental setup, i.e., a listening experiment in which we investigate participants' ability to hear individual naturally sounding speech signals in a mixture with reduced spatial and speaker cues. In particular, we present the audio signal to the listener as a *monaural mixture* of two different narratives uttered by a *speech synthesizer*

(TTS project at AT&T Labs Research [17]), using the same virtual speaker. In this way, we eliminate cues to separation related to different spatial locations of the sound sources, different accents, different genders of the speakers etc. However, we still want the speech to sound natural, hence, the speech synthesizer does produce prosodic speech which also provides a cue to stream separation and tracking [18]. It is known that the introduction of this kind of voice modifications effects detectability [19], however, we expect these effects to be reduced compared to conventional human speech. To further reduce energetic masking, we equalized the total energy of the mixed signals.

In order to reduce basic semantic masking effects we opted for narratives with little expected interest to the listeners. In particular we chose as excerpts from neutral texts used in preparation for the TOEFL (Test of English as a Foreign Language) test [20]. These texts are more coherent and naturally sounding than the short command sentences used in [16].

We have recently proposed a computational mechanism for task driven top-down attention based on a generative statistical model of inputs, corresponding to a 'gist' of the scene and to potential elements for attention, and task labels corresponding to possible actions chosen based on gist and attended inputs ([5], [6]). The notion of gist refers to unspecific inputs generated in part by bottom-up attention [21]. This model can be used to make predictions about the influence of confounders on task labeling performance, in both strongly task dependent attention and under weakened task influence. Thus we have designed experiments so that they test both of these cases. In the first set of experiments (we call it *undirected attention* (UNDIR)), the subjects do not have any specific task rather than simply aim to hear as many words as possible, thus they may follow any of the two narratives, while, in the second set of experiments (*directed attention* (DIR)), they are asked to focus on just one of the two narratives (the choice about which of them is left to the subject).

To test the relative influence of top-down and bottom-up information flow on attention and masking we estimate new overlap scores defined in this papers and based on the so-called ideal binary mask (IBM) [22]. An IBM consists of zeros and ones where ones represent the powerful parts of the target audio signal compared to an interference audio signal. IBMs have been shown to improve human speech intelligibility when applied to noisy speech signals. Subjects have been exposed to the re-synthesized speech signals from the IBM-gated (segregated) signal and they recognized words quite well even for a signal-to-noise-ratio (SNR) as low as -60 dB which corresponds to pure noise ([23], [24]). In addition,the features obtained from IBMs have worked successfully for an automatic speech recognition (ASR) application [8].

The influence of masking is measured as the correlation between the number of times specific words are heard (WOH) and the relative overlap. IBM control parameters are first chosen taking as references previous works on speech intelligibility [24] and ASR [8] as a pre-analysis step. Then, we optimize those parameters to have the most negative correlation. In particular, we analyze local criteria (LC), the window length (winLength) and number of frequency channels (numChan).

The paper is organized as follows; first we discuss the proposed top-down attention model and the experiments designed to investigate the role of attention in the hard cocktail party problem, we present simulation results of the model and experimental results, and finally give our interpretation of the findings in the discussion section.

II. Methods

A. Attention Model

By means of movement and information processing the brain actively selects its input. In the broadest sense attention is the mechanism by which the brain selects relevant input. Bottom-up attention is typically driven by statistical novelty, i.e., we attend to the un-expected, while top-down attention select input that is relevant to a given task. While the latter definition is in broad consensus, there has been remarkably few attempts to formalize top-down attention as a statistical problem. In [5] we model the top-down attention as a decision problem based on incomplete information and analyze which feature to measure next in a classification problem.

Attention is implemented in a statistical model as the selection of additional input features based on an initial subset of features representing the 'gist'. We attend to features that reduce confusion at the models' output level, i.e., features that have high expected information given the 'gist'.

We represent the task by probability distribution over a set of C 'actions' or classes indexed by the discrete variable c, $(c = 1, ..., C)$. Initially, for decision making we have access to the gist, a vectorial observation $\boldsymbol{x}$ with components x_i, $i = 1, ..., I$. The second step concerns an additional measurement z_j which is obtained by attending to a specific channel j, chosen among the set of missing features $\boldsymbol{z}$ with components $z_1, ..., z_J$. The joint probability of the classes and all features observed and missing is $p(c, \boldsymbol{x}, \boldsymbol{z})$. Attention is based on the information available in $\boldsymbol{x}$, giving us the pre-attention posterior probabilities for the task

$$\begin{aligned} p(c|\boldsymbol{x}) &= \int p(c, \boldsymbol{z}|\boldsymbol{x}) d\boldsymbol{z} \\ &= \frac{\int p(c, \boldsymbol{x}, \boldsymbol{z}) d\boldsymbol{z}}{\sum_{c=1}^{C} \int p(c, \boldsymbol{x}, \boldsymbol{z}) d\boldsymbol{z}}. \end{aligned} \tag{1}$$

Using the top down attention mechanism we will select an additional feature $\boldsymbol{z}_j$, which will result in the distribution

$$\begin{aligned} p(c|\boldsymbol{x}, z_j) &= \sum_{c=1}^{C} \int p(c, \boldsymbol{z}|\boldsymbol{x}) \prod_{i \neq j} dz_i \\ &= \frac{\int p(c, \boldsymbol{x}, \boldsymbol{z}) \prod_{i \neq j} dz_i}{\sum_{c=1}^{C} \int p(c, \boldsymbol{x}, \boldsymbol{z}) \prod_{i \neq j} dz_i} \end{aligned} \tag{2}$$

The information value of this choice is given as the difference in confusion (entropy) before and after the attended measurement, which will depend on the particular outcome of the sequential measurement, z_j,

$$\begin{aligned} \Delta S_j(\boldsymbol{x}, z_j) &= \sum_{c=1}^{C} \log p(c|\boldsymbol{x}, z_j) p(c|\boldsymbol{x}, z_j) \\ &- \sum_{c=1}^{C} \int \log p(c, \boldsymbol{z}|\boldsymbol{x}) p(c, \boldsymbol{z}|\boldsymbol{x}) d\boldsymbol{z} \end{aligned} \tag{3}$$

As z_j is unknown before attending, we base the choice on the expected gain *expected information gain* of measuring the value of

feature j,

$$\begin{aligned} G_j(\boldsymbol{x}) &\equiv \int \Delta S_j(\boldsymbol{x}, z_j) p(z_j|\boldsymbol{x}) dz_j \\ &= \sum_{c=1}^{C} \int \log p(c|\boldsymbol{x}, z_j) p(c, z_j|\boldsymbol{x}) dz_j \\ &\quad - \sum_{c=1}^{C} \int \log p(c, \boldsymbol{z}|\boldsymbol{x}) p(c, \boldsymbol{z}|\boldsymbol{x}) d\boldsymbol{z}. \end{aligned} \quad (4)$$

The Gaussian Discrete mixture model (GDMM) is a generative model of the joint distribution, see e.g., [25]

$$p(c, \boldsymbol{x}, \boldsymbol{z}) = \sum_{k=1}^{K} p(k) p(c|k) p(\boldsymbol{x}, \boldsymbol{z}|k) \quad (5)$$

where K is the number of components, $p(k)$ are component probabilities, $p(c|k)$ is a $C \times K$ probability table, and $p(\boldsymbol{x}, \boldsymbol{z}|k)$ are K Gaussian pdfs. The GDMM is convenient as both conditioning and marginalization are computationally tractable. We choose a generative representation to allow for modeling of input dependencies which is necessary in order to make inference about missing features. Maximum likelihood parameter estimation in the GDMM leads to a straightforward generalization of expectation maximization algorithm for conventional mixtures.

If we introduce the GDMM in the information gain expression we obtain

$$\begin{aligned} G_j(\boldsymbol{x}) &= \sum_{c=1}^{C} \sum_{k=1}^{K} p(c|k) p(k|\boldsymbol{x}) \times \\ &\qquad \int \log \left[p(c, \boldsymbol{x}, z_j) \right] p(z_j|\boldsymbol{x}, k) dz_j \\ &- \sum_{k=1}^{K} p(k|\boldsymbol{x}) \int \log \left[p(\boldsymbol{x}, z_j) \right] p(z_j|\boldsymbol{x}, k) dz_j \\ &+ \text{const.} \end{aligned} \quad (6)$$

where $p(c, \boldsymbol{x}, z_j) = \sum_{k=1}^{K} p(k) p(c|k) p(\boldsymbol{x}, z_j|k)$ and $p(\boldsymbol{x}, z_j) = \sum_{k=1}^{K} p(k) p(\boldsymbol{x}, z_j|k)$. Thus, computing G for all I features amounts to computing $Q = I * (C + 1) * K$ one-dimensional integrals over Gaussian measures $p(z_j|\boldsymbol{x}, k) = \mathcal{N}(\mu_j(\boldsymbol{x}, k), \sigma_j^2(\boldsymbol{x}, k))$ with

$$\begin{aligned} \mu_j(\boldsymbol{x}, k) &= \mu_{j,k} + \boldsymbol{\Sigma}_{z_j, \boldsymbol{x}, k} \boldsymbol{\Sigma}_{\boldsymbol{x}, \boldsymbol{x}, k}^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_{\boldsymbol{x}, k}) \\ \sigma_j^2(\boldsymbol{x}, k) &= \sigma_{j,k}^2 - \boldsymbol{\Sigma}_{z_j, \boldsymbol{x}, k} \boldsymbol{\Sigma}_{\boldsymbol{x}, \boldsymbol{x}, k}^{-1} \boldsymbol{\Sigma}_{\boldsymbol{x}, z_j, k}. \end{aligned} \quad (7)$$

In these expressions $\mu_{j,k}, \sigma_{j,k}^2$ are the mean and variance of the jth feature in the kth component, while $\boldsymbol{\Sigma}_{a,b,k}$ is the part of the covariance matrix of the k component corresponding to variable sets a, b.

More details and further references are given in [5], where the attention model was validated on four benchmark classification problems and shown to outperform a 'random' attention alternative.

To simulate strong and weak top-down attention we augment the model by smoothing the label-component table $p(c|k) \rightarrow p(c|k, \beta) = p(c|k)^{\beta} / \sum_c p(c|k)^{\beta}$ and by letting the attention selection be stochastic based on the expected gains, i.e., we select attention using the induced probability distribution,

$$P(j) = \frac{\exp(\gamma G_j)}{\sum_{j'} \exp(\gamma G_{j'})} \quad (8)$$

Task-driven top-down attention as in [5] is obtained when $\beta = 1, \gamma = \infty$. In this work we use $\beta = 0.2, \gamma = 0.33$.

We challenge the strong and weak top-down attention models by a simulated cocktail party by confounding the input of test data for a $C = 2$ environment. In particular we mix for each pattern a fraction f (mixing fraction) of the input features with a randomly chosen input feature vector from the opposite class (confounder):

$$overlapped\ signal = (1 - f) * input\ signal + f * confounder$$

The $C = 2$ simulated decision problem is based on a four component Gaussian mixture, the resulting configuration is first established in two dimensions and resembles the well-known XOR-problem, hence can not be separated linearly. The two dimensional input space is augmented by six noise dimensions $SNR \approx 1$, so that the total input dimension is eight. In the attention experiments one signal dimension and one noise dimension is provided as 'gist'.

B. Behavioral Experiments

While it is not possible to directly read out the informations flows in the human brain while solving a difficult speech separation task, some insight can be obtained by observing the macroscopic behavior. Here we design a behavioral experiment inspired by the pioneering work of Cherry [1]. Basically, we design a hard cocktail party problem by reducing conventional auditory cues as described above, leaving only high-level cognitive cues such as semantic and context representation in narratives. Cherry alluded to these high-level representations as what he called word *transition probabilities*. Our hypothesis is that these representations precisely are subject to top-down attention and should be task dependent, while the more basic cues could be more automatic and operate beyond conscious control.

Subjects are presented with two different narratives combined in a mono audio file with a headphone. The stories are generated by a speech synthesizer (TTS project at AT&T Labs Research [17]), using the same virtual speaker.

We recruited twelve participants among master students, PhD students and post-docs from the Technical University of Denmark. Two participants were not able to accomplish any of the experiments and were excluded.

We perform two different types of experiments which we will refer to as *undirected attention* and *directed attention* experiments. In the first case the listener is free to follow either narrative. In the directed attention case, participants are asked to follow one narrative story at their own discretion. At the end of an audio presentation, a list of terms is presented and they are asked to check which terms they have heard. The list contains words which are in the narratives and words which are not, but are related to the content.

Cherry made his participants listen as many times as the they wanted; we perform three different trials (4 people for each trial). In the first and in the second, we make them listen just once and then we ask for the words. They repeat the same process three times. In the third, we make them listen twice before presented with the term lists.

The words in the list can appear various numbers of times in the narratives, but we balance this number, in total, for both tracks. In particular, the total number of occurrences of all the words we ask for is the same in both stories. Moreover, we aimed to balance the frequency of each word in the list; which means, for example, that if there are two words appearing in the list, respectively, three and five times in the first story, there are also two words appearing three and five times in the second one. The list of words contains 48 words: 24

of these are truly present in the audio signal and each story contains half of them.

We use two different narratives for each experiment, making small changes (removing or adding sentences or words, switching the order of some sentences or words) to the original texts. This is necessary to have stories with the same length and in which pauses are synchronized as much as possible, to avoid the so called 'listening in the gaps' effect, described by Bregman [9] and by Bronkhorst [12].

C. Audio Analysis

The basis for our audio analysis is the *ideal binary masks* (IBMs) which we use to measure the cross channel interference in terms of temporal and spectral overlap. It is obtained by comparing the spectrogram of a target sound signal to that of the interference signal and to keep only the strong time-frequency regions of it. More specifically, its value is one when the target is stronger than the interference for a local criterion (LC), and zero elsewhere. The time-frequency (T-F) representation is obtained by using the model of the human cochlea as the basis for data representation [26]. If $T(t, f)$ and $I(t, f)$ denote the target and interference time-frequency magnitude, then the IBM is defined as

$$\mathrm{IBM}(t, f) = \begin{cases} 1, & \text{if } T(t, f) - I(t, f) > LC \\ 0, & \text{otherwise} \end{cases} \quad (9)$$

In Figure 1, we show an example of an IBM obtained with a sample sound from one of the stories (the sound relative to the word 'navigate') compared to a speech shaped noise (SSN) as the interference signal. The most energetic parts of the target signal are kept. We measure the spectral and temporal overlap between two

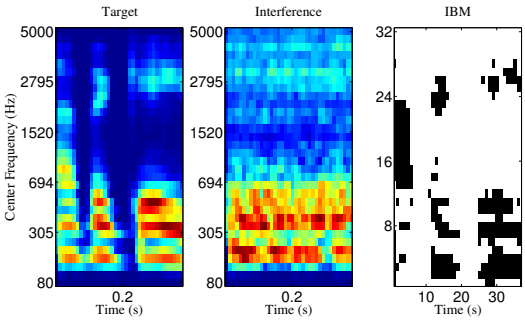


Fig. 1. The spectrograms of a target sound signal, the interference signal (SSN) and the resultant IBM (SNR=0 dB, LC=-4 dB, windows length=20 ms (50% overlap), frequency channels=32, frequency bins are not equally spaced, gammatone filtering is used, 1 = black 0 = white)

sound signals, specifically between a word in one of the narratives pronounced by the speech synthesizer and the corresponding part in the second story. We define the temporal overlap between them as a percentage of the whole duration without silence in the time domain. We use IBMs of the sound signals as mentioned before and we compress both IBMs over frequency. For a time slot, we assign one if there is at least one one on the mask, otherwise zero where zeros are considered as silence. Then, the temporal overlap is simply the overlap of ones on the masks (see Figure 2).

The spectral overlap is defined similarly based on co-occurrence of signals in the time-frequency bins. Once we have IBMs for both sound signals, we simply compute the percentage of the overlapping ones on both masks over the whole time-frequency bins (see Figure 2).

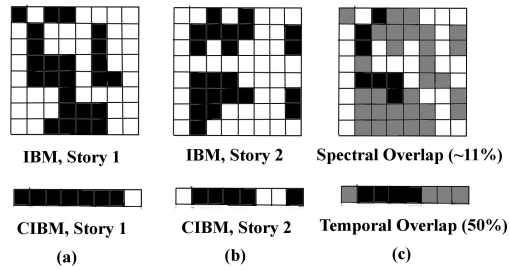


Fig. 2. The illustrations of temporal and spectral overlap definitions, the bins represent time-frequency regions of an IBM (frequency bins are not equally spaced, gammatone filtering is used). Only black regions represent overlapped parts on (c).

Based on the temporal and spectral overlaps for words in both stories, we explore the correlation between the overlaps and the number of times the words were heard by the subjects for both directed and undirected cases.

The overlap values depend on parameters that change the resultant IBMs, while the number of times the words were heard is fixed. In particular, for each word, the number we consider in the analysis is computed averaging out the total number of times the word was heard (in all the experiments, in all the trials, by all the subjects) against the number of times the same word occurs in the stories. IBM parameter values are first chosen taking as references previous works on speech intelligibility [24] and on ASR [8] as a pre-analysis step. In [24], subjects listened to IBM-gated (multiplying the spectrogram of a noisy-speech with IBM of it, and resynthesizing in time domain) and for a range of IBM parameter values, the best speech intelligibility results(recognizing which word is pronounced) are obtained. In [8], the best performance for an ASR system is obtained again with same range of values. However, referenced to those studies, even if those parameters are expected to result in overlaps closer to what humans perceive, they are not necessarily optimal to investigate the correlation between overlap and word detection rates. Therefore we optimize the IBM parameters including the local criteria (LC) with fixed SNR, the windows length and the number of frequency channels to gain the most negative correlation. We keep other two parameters constant while optimizing one. With the optimized values, we apply a permutation test with 10000 resamples, at 5% significance level, to validate the results.

III. Experimental Results

First, we set up a simulation experiment with the top-down attention model emulating the Cherry experiment, as described above. For a range of mixing fractions $f \in [0.0\ 0.2]$ we measure the resulting error rates for the models using the attention mechanism to select an additional feature among the six remaining after having provided the 2-dimensional gist. The strong and weak top-down attention response is shown in figure 3. The rates are represented as relative excess errors: $[E(f) - E(0)]/(B - E(0))$, where $B = 0.5$ is the baseline error rate. The error rate of the strong task attention model is $E_{\mathrm{DIR}}(0) = 0.08$ while the error rate of the weak task attention is $E_{\mathrm{UNDIR}}(0) = 0.23$. The experiment indicates that strong top-down attention model (DIR) is less sensitive to the confounding

mixture than the weak attention model, hence it will make more informed decisions in the cocktail party.

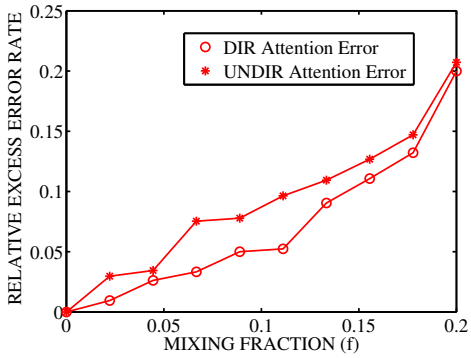


Fig. 3. The resulting error rates for the models using the attention mechanism to select an additional feature among the six remaining after have provided the 2-dimensional gist for a range of mixing fractions $f \in [0.0\ 0.2]$.

The behavioural experiments are carried out using a GUI implemented in Java, while the results are analysed using MATLAB. The word boundaries are determined manually to be more precise (The limited number of words enables us to do so). The sampling frequency of the audio signals is 16 kHz. We use gammatone filters which is a commonly used model for auditory filters in the auditory system to obtain IBMs.

Figures 4 and 5 show the temporal and spectral overlaps for each word, for UNDIR and DIR cases respectively, using non-optimized parameters from [24], [8]. We observe that the correlation between overlaps (temporal and spectral) and rate of heard words are -0.35 and -0.31 respectively. While, for DIR case, we find positive correlations of 0.23 for temporal and 0.34 for spectral. We next optimize the

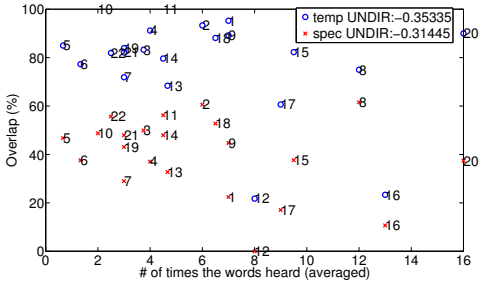


Fig. 4. Temporal and spectral overlap versus the averaged number of times the words heard (WOH) for UNDIR case, and the correlation between them shown on the legend

three IBM parameters, LC, WinLength and NumChan, to produce the most negative correlation between overlaps and words detected. The resulting figures show that optimal LC values are around -10dB for all cases except for the spectral overlap in the UNDIR case, which is -14dB (see Figure 6). We also conclude that 20ms is the optimal value for the windows length for all cases (see Figure 7). We see that for the spectral overlap in the DIR case, the correlation values for WinLength

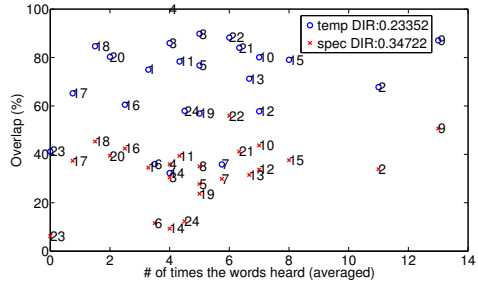


Fig. 5. Temporal and spectral overlap versus the averaged number of times the words (WOH) heard for DIR case, and the correlation between them shown on the legend

greater than 20ms are not present. This is due to the fact that with the high optimal value found previously, the resultant IBMs were all zeros (we did not try to play with the values, because it is already hard to find significant results for DIR case). Finally, we observe that the optimal values for number of frequency channels is 16 and 32 (see Figure 8). Using optimal IBM parameters for each case

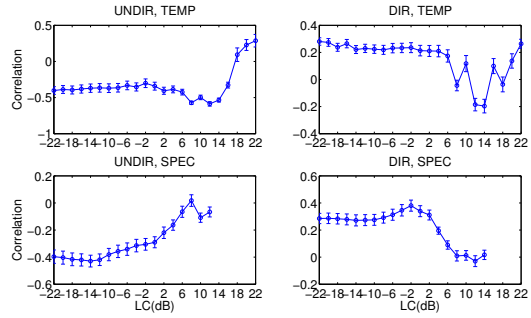


Fig. 6. Correlation for different LC values, WinLength = 20ms and NumChan = 32. Left to right: Undirected and Directed. Top to bottom: Temporal and Spectral.

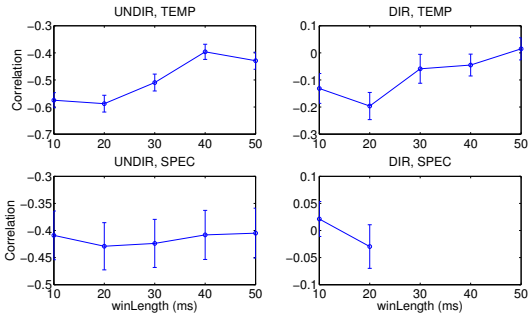


Fig. 7. Correlation for different WinLength values, optimal LC for each case and NumChan = 32. Left to right: Undirected and Directed. Top to bottom: Temporal and Spectral.)

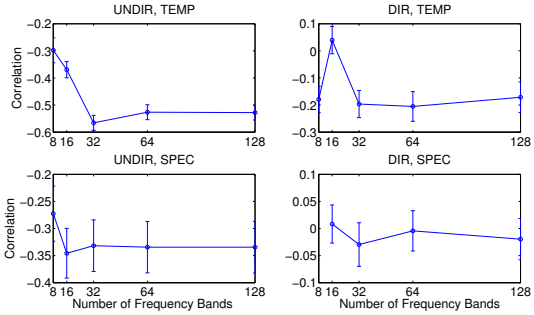


Fig. 8. Correlation for different NumChan values, optimal LC and WinLength for each case. Left to right: Undirected and Directed. Top to bottom: Temporal and Spectral.)

(UNDIR,DIR, temporal and spectral) we obtain similar results. The correlations between overlaps (temporal and spectral) and WOH are -0.59 and -0.43 (more negative than not-optimized case) respectively for UNDIR case. However, for DIR case, they are -0.20 for temporal and -0.03 for spectral. Even if the results for DIR case also are more negative than not-optimized case , they are evidently less than those of UNDIR case (almost no correlation for DIR spectral case). Finaly, we apply the permutation test to these data, as mentioned in the section II-C. In both spectral and temporal overlaps, for UNDIR experiments, under the 5% significance level, the null hypothesis that the data is uncorrelated is rejected (spectral = 3.1% and temporal = 0.07%). In the DIR experiments the null hypothesis is accepted, indicating the influence of masking is compensated by a more detailed model.

A. Discussion and Conclusion

Based on our recent top-down attention model we can simulate the cocktail party effect. We found that the top-down attention model showed less sensitivity to the amount of the confounding overlap, than the weak attention model. This indicates that the top-down mechanism can assist to compensate for structured noise.

In the 'hard cocktail party' behavioral experiment we found significant negative correlations between overlaps of two concurrent sounds and speech intelligibility for the data collected in the undirected attention experiments (UNDIR, no task). While in the directed attention experiments (DIR, task-driven) we accepted the no-correlation null hypothesis, even after careful optimization for correlation, a finding well-aligned with the simulation result.

We conclude that the relation between energetic masking and speech intelligibility is modulated by the presence of a task, hence top-down controlled attention. Based on our top-down attention model we expect this to be a special case of a more general phenomenon, namely that the top-down knowledge can enhance pattern recognition by compensating for noise and the presence of confounders.

REFERENCES

[1] E. C. Cherry, "Some experiments on the recognition of speech, with one and two ears," *Journal pf Acoustic Society of America*, vol. 25, pp. 975–979, 1953.

[2] M. McKeown, L. K. Hansen, and T. J. Sejnowski, "Independent component analysis for fmri: What is signal and what is noise?" *Current Opinion in Neurobiology*, vol. 13, no. 5, pp. 620–629, 2003.

[3] A. J. Viterbi, *CDMA: principles of spread spectrum communication*. Redwood City, CA, USA: Addison Wesley Longman Publishing Co., Inc., 1995.

[4] A. Hyvarinen, J. Karhunen, and E. Oja, *Independent component analysis*. J. Wiley, 2001.

[5] L. Hansen, S. Karadogan, and L. Marchegiani, "What to measure next to improve decision making? On top-down task driven feature saliency," in *SSCI, IEEE Symposium on Computational Intelligence, Paris, France. CCMB Cognitive Algorithms, Mind, and Brain*, 2011, pp. 86–87.

[6] S. Karadogan, L. Marchegiani, J. Larsen, and L. Hansen, "Top-down attention with features missing at random," in *IEEE International Workshop on Machine Learning For Signal Processing*, 2011, submitted.

[7] S. Choi, H. Hong, H. Glotin, and F. Berthommier, "Multichannel signal separation for cocktail party speech recognition: A dynamic recurrent network," *Neurocomputing*, vol. 49, no. 1-4, pp. 299–314, 2002.

[8] S. Karadogan, J. Larsen, M. Pedersen, and J. Boldt, "Robust isolated speech recognition using binary masks," *European Signal Processing Conference, EUSIPCO*, 2010.

[9] A. Bregman, *Auditory scene analysis: The perceptual organization of sound*. The MIT Press, 1994.

[10] B. Moore and H. Gockel, "Factors influencing sequential stream segregation," *Acta Acustica United with Acustica*, vol. 88, no. 3, pp. 320–333, 2002.

[11] R. Drullman and A. Bronkhorst, "Multichannel speech intelligibility and talker recognition using monaural, binaural, and three-dimensional auditory presentation," *The Journal of the Acoustical Society of America*, vol. 107, p. 2224, 2000.

[12] A. Bronkhorst, "The cocktail party phenomenon: A review of research on speech intelligibility in multiple-talker conditions," *Acta Acustica united with Acustica*, vol. 86, no. 1, pp. 117–128, 2000.

[13] M. Bee and C. Micheyl, "The cocktail party problem: What is it? how can it be solved? and why should animal behaviorists study it?." *Journal of Comparative Psychology*, vol. 122, no. 3, p. 235, 2008.

[14] R. Cusack, J. Decks, G. Aikman, and R. Carlyon, "Effects of location, frequency region, and time course of selective attention on auditory scene analysis." *Journal of Experimental Psychology: Human Perception and Performance*, vol. 30, no. 4, p. 643, 2004.

[15] R. Carlyon, R. Cusack, J. Foxton, and I. Robertson, "Effects of attention and unilateral neglect on auditory stream segregation." *Journal of Experimental Psychology: Human Perception and Performance*, vol. 27, no. 1, p. 115, 2001.

[16] D. Brungart, "Informational and energetic masking effects in the perception of two simultaneous talkers," *The Journal of the Acoustical Society of America*, vol. 109, p. 1101, 2001.

[17] M. Beutnagel, A. Conkie, J. Schroeter, Y. Stylianou, and A. Syrdal, "The at&t next-gen tts system," in *Joint Meeting of ASA, EAA, and DAGA*. Citeseer, 1999, pp. 18–24.

[18] A. Syrdal and Y. Kim, "Dialog speech acts and prosody: Considerations for tts," in *Proceedings of Speech Prosody*, 2008, pp. 661–665.

[19] M. Jilka, A. Syrdal, A. Conkie, and D. Kapilow, "Effects on tts quality of methods of realizing natural prosodic variations," in *Proc. ICPhS*, 2003.

[20] D. Phillips, *Preparation Course for the TOEFL: Next Generation IBT*. Longman, 2006.

[21] A. Oliva and A. Torralba, "Building the gist of a scene: the role of global image features in recognition," in *Progress in Brain Research*, 2006, p. 2006.

[22] D. Wang, "On ideal binary mask as the computational goal of auditory scene analysis," *Speech separation by humans and machines*, pp. 181–197, 2005.

[23] D. Wang, U. Kjems, M. Pedersen, J. Boldt, and T. Lunner, "Speech perception of noise with binary gains," *The Journal of the Acoustical Society of America*, vol. 124, pp. 2303–2307, 2008.

[24] U. Kjems, J. Boldt, M. Pedersen, T. Lunner, and D. Wang, "Role of mask pattern in intelligibility of ideal binary-masked noisy speech," *The Journal of the Acoustical Society of America*, pp. 1415–1426, 2009.

[25] L. K. Hansen, S. Sigurdsson, T. Kolenda, F. Å. Nielsen, U. Kjems, and J. Larsen, "Modeling text with generalizable gaussian mixtures," in *ICASSP International conference on acoustics, speech and signal processing*, vol. 4, 2000, pp. 3494–3497.

[26] R. Lyon, "A computational model of filtering, detection, and compression in the cochlea," in *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP'82.*, vol. 7, 1982, pp. 1282–1285.

Appendix F

Dimensional Emotion Recogniton from Speech Combining Semantic and Acoustic Features

---

S.G. Karadogan, J. Larsen. Dimensional Emotion Recogniton from Speech Combining Semantic and Acoustic Features. *IEEE Transaction on Affective Computing*, submitted, 2012.

Dimensional Emotion Recognition from Speech Combining Semantic and Acoustic Features

Seliz G. Karadogan, *Member, IEEE,*
Jan Larsen, *Senior Member, IEEE*

Abstract—Most of the previous studies on speech emotion recognition focused on the 'the categorical approach' classifying basic emotions (such as happiness and anger). However, the focus has been shifting towards 'the dimensional approach' in which researchers argue that emotional states are not independent from one another but related in a systematic manner. There has been also effort on integration of linguistic and acoustic information. This paper investigates novel approaches to dimensional speech emotion recognition using both audio and textual information and the role of acoustic and semantic features in arousal and valence dimensions. We design a new corpus that consists of 59 short movie clips with audio and text, rated by human subjects in arousal and valence (A-V) dimensions. We use the Affective Norms for English Words (ANEW), that are rated also in A-V dimensions by humans, as emotional keywords in semantic part. The semantic and acoustic features are processed separately, and finally, the results of both parts are combined. We do not only show that integration of semantic and acoustic features improve the performance, but also show that valence is better estimated using semantic features while arousal is better estimated using acoustic features.

Index Terms—emotion recognition, arousal, valance, semantics, acoustics.

1 Introduction

The fact that speech is considered as a natural and effective way of communication between humans, directed researchers devoting much work into speech analysis, considering it as an efficient method for human computer interaction (HCI) as well. The speech does not only carry explicit messages but also various sources of information about the speaker such as gender, age, physiological and emotional states. Emotions are fundamental parts of those implicit messages and for better HCI systems recent research has focused on the recognition of emotional speech. It has been shown that it is crucial for communication systems to recognize emotional states of humans [1].

This paper is an extension of our work in [2]. We design a speech emotion recognition model with a dimensional approach that integrates the acoustic and semantic features, using the Affective Norms for English Words (ANEW [3]) as reference emotional keywords. We design our own emotional database using short audio clips from English movies. The feelings expressed in audio clips are rated by human observers on valence and arousal dimensions. We use the same figures for rating (self-assessment manikin (SAM) [4]) and similar instructions used in ANEW work for designing the user interface to gather ratings, to coincide with the rating procedure of ANEW work that is a reference to our work. We extract and combine hundreds of acoustic features, use the correlation based feature selection (CFS [5]) method to reduce the number of features and use support vector regression (SVR) [6] as the learning model, using the openEar toolkit explained in [7]. To extract semantic features, we use Latent Semantic Analysis (LSA [8]) to find similarities of each term in the text of each clip to ANEW words (word by word similarities) and evaluate the dimensional ratings of each word combining those similarities with A-V values of ANEW words. We estimate the ratings of each clip using acoustic and semantic features separately, and finally, we combine the results of both parts.

1.1 Background

Darwin in [9] suggested that emotions are an evolved communication mean and therefore should prime behaviors that enables others to conceive one's emotional state. It is possible to reveal one's emotional state from vocal characteristics, facial displays, and whole-body behaviors [10]. Although, intuitively it is an easy task to determine whether someone is experiencing a particular emotion, how humans perceive emotions is one of the most important questions still in search in affective science. There are two main emotional modelings approaches: categorical and dimensional approaches.

• *The authors are with the Department of Informatics and Mathematical Modeling, Technical University of Denmark, Copenhagen, Kgs. Lyngby, 2800.*
E-mail: seka@imm.dtu.dk, jl@imm.dtu.dk

The categorical approach is based on classification of discrete emotions such as happiness, sadness and surprise that are hard-wired in human brain and recognized universally [11]. The classification is usually based on 'basic' emotions (anger, fear, disgust, happiness, sadness, and surprise) which have been showed by Ekman in [12], following Darwin's analysis, to be recognized cross-culturally by facial behaviors. Ekman in [13], explains the 2 meanings of the adjective 'basic' in this context. The first meaning of 'basic' indicates emotions differ from each other not only in expression but also in other aspects such as appraisal, probable behavioral response, physiology, etc. The second meaning instead indicates that emotions evolved dealing with fundamental life-tasks.

In the dimensional approach, researchers argue that emotional states are not independent from one another but related in a systematic manner. The most commonly assumed dimensions are valence, arousal, approach-avoidance and dominance-submission [14], [15], [16]. Valence and arousal dimensions have been shown to cover the majority of affect variability and been widely used [11], [17]. The arousal dimension represents how excited the emotion is or how much energy is required to express the feeling. Feelings with high arousal induce some physical changes in the body such as increased heart rate, higher blood pressure and greater sub-glottal pressure resulting in change in speech as well such as making it louder, faster and have higher average pitch etc. The valence dimension refers to how positive or negative the emotion is, ranging from pleasant to unpleasant. However, it has not been shown yet how or if the acoustics features correlate with the valence dimension [18]. Even if there exists some works done on dimensional automatic emotion recognition [19], [20], it is still in its infancy as also stated in [11]. Dominance-submission is a continuum ranging from feeling influential and in control to the opposite extreme feeling of lack of control or influence. Meanwhile, approach motivation is characterized by tendencies to approach stimuli as in excitement or interest and avoidance motivation is characterized by tendencies to avoid stimuli as in anxiety or shame.

Researchers still did not conclude which dimensional scheme should be used to measure emotion. However, in [10], it is claimed that different measures of emotion are sensitive to different dimensions. They concluded that self-report of one's current emotional state, autonomous nervous system (ANS) measure in which skin conductance level, heart rate, etc. are measured, are sensitive to both valence and arousal dimensions. The startle response magnitude that measures startle in response to a sudden, intense stimulus including tensing of the neck and back muscles and an eye blink and Electroencephalography (EEG) are sensitive to approach and avoidance motivations. Vocal characteristics such as pitch, amplitude, etc. is sensitive to arousal dimension, while observer ratings of facial expressions is sensitive to valence dimension.

Although emotions have been conceptualized in both dimensional and categorical terms, some recent works support the dimensional perspective [10]. In [10], they reviewed that it is more difficult to establish emotion specificity in the domains of ANS activity, affect-modulated startle responses, and vocal characteristics and they concluded dimensions capture most of variance of emotional responses. Meanwhile, dimensional and categorical perspectives can be combined to some extent by describing discrete emotions in terms of multiple dimensions (e.g., anger negative valence, high arousal) [10], [21].

1.2 Related Work

The extraction of suitable features is one of the challenges in speech emotion recognition task [18]. Semantic and acoustic features, which are based on what and how it is said respectively, can be gathered from speech. Most effort has been made to the use of acoustic features for speech emotion recognition [22], [23]. The typical and mostly used features are the pitch, the formants, the short-term energy, the mel-frequency cepstral coefficients (MFCC), and the Teager energy operator [24]. Many acoustic features have been explored, however, there is still no standard group of features defined, which efficiently characterizes the emotional content, independent of the speaker and the lexical content.

The second challenge is dealing with the available large number of acoustic features extracted from speech. Hundreds of features could be obtained which leads redundant information and a computational burden. In many cases, feature reduction and selection methods are applied to deal with huge number of features [25], [26], [27], [28], [29]. In [28], information gain as feature reduction method, in [25], Sequential Forward Floating Search (SFFS) and in [26], forward selection (FS) and in [29], correlation-based feature selection (CBFS) as feature selection methods have been used.

The choice of a learning model is important as well once features are extracted. Many models have been used for this task, such as hidden Markov model (HMM) [22], [30], support vector machines (SVM) [25], [31], [32], K-nearest-neighbor (KNN) [31], Gaussian mixture models (GMM) [33] and neural networks (NN) [34], [35], but there is yet no agreement on the most suitable one.

Recently, much effort has been put on the integration of acoustic and semantic information of speech [20], [36], [37]. It has been shown in recent works that combining acoustic and semantic features improves emotion recognition results [20], [37]. Bag-of-words (BOW) features, where each term within a vocabulary is represented by a feature modeled by the

term's occurrence frequency within the phrase, and part-of-speech (POS) features, where each phrase is represented by grammatical tags (verbs, nouns, etc.), have been shown to be useful linguistic features for speech emotion recognition task [20], [27], [38]. The vocabulary is often limited to predefined emotional keywords including words like 'happy', 'sad' and 'depressed' [37]. Nevertheless, not only keywords but also general terms may carry the emotional content [39]. The sentences 'I passed the exam' and 'I am happy to have passed the exam' carry similar emotions, but the former lacks the keyword 'happy'. Semantic language analysis, which is the study of 'meaning' and is a subfield of linguistics, has been investigated addressing to this problem [40]. Latent semantics analysis (LSA) is an indexing method that is based on the principle that words occurring in similar contexts are also similar in meanings [8]. It has often been used for text-emotion recognition task [41], [42]. In [40], LSA has also been used for the emotional analysis of songs, using the lyrics, where they measure the similarities of lyrics as a whole to the emotional keywords defined. In [43], they work on detecting emotional state of a child in a conversational computer game and they also make use of LSA to find the similarities (they call the similarity found as 'emotional distance') between test and training speech utterances.

There have been studies in affective science about how to evaluate each word in the emotional dimensions which could have an important impact on semantic analysis part of speech emotion recognition task. The affective norms for English words (ANEW) in [3] introduced by Bradley and Lang in 1999, includes a set of normative emotional ratings for 1034 English words, in arousal, valence and dominance dimensions. It has been developed to provide researchers with the standardized materials in emotion studies. The self-assessment manikin (SAM), an affective rating systems designed by Lang in 1990 [4], is used to assess the three affect dimensions which are valence, arousal and dominance. The graphic SAM figures has nine values, from low to high and neutral in the middle, comprising bipolar scales in each dimension. ANEW has been widely used for emotion analysis purposes by researchers and has also been adapted for different languages [44], [45], [46], yet, how to rate and assess emotions is still questioned.

Another challenge for speech emotion recognition task is the acquisition of an appropriate database. There are many aspects to be considered while trying to choose a database, such as the language, the scope (emotion analysis or recognition), subjects and observers (adults or children), naturalness (acted or natural), the balance in phrases (the number of phrases per emotion, phrase length, etc.), the emotion model (categorical or dimensional), the assessment type (which emotions for the categorical, continuous or quantized for dimensional, etc.) and the duration of phrases or dialogs, etc. There are a number of available databases developed well, however, it is usually hard to find one that is convenient in all those aspects. Therefore, some prefer to design their own emotional speech databases that apply to their research aims [37], [47].

1.3 Paper Outline

In the following section, we explain the design process of the emotional speech database we created, the EMOV database and we analyze the data obtained. In section 3, we give an overview of the modeling framework, and explain the analysis process of acoustics and semantics parts separately and of the integration of these parts. Later in section 4, we describe the experiments and give the results for acoustics, semantics parts and for the combination of both. Finally, we discuss the results of our study in section 5 and make a short conclusion in section 6.

2 Database Design

2.1 Design

Two important parts in our work is that we make semantic analysis on text part of the speech and use ANEW words as reference keywords and these two parts bring specific requirements for the database we need. The requirements for the database we needed are the following:

- Language: English
- Emotion Model: Dimensional model (including valence and arousal)
- Response format: Similar to the ANEW format as much as possible
- Annotators: Adults
- Content : Speech including words with emotional content (not only neutral words like date, time etc.)

Unfortunately, we were not able to find a database that matched our requirements. There are only a few emotional speech databases that are publicly available for use [18], [48][1]. Most of these databases are intended for categorical approach, thus the audio samples are rated for discrete emotions such as anger, happiness, fear etc. [49], [50]. Recently, there have been attempts for acquisition of emotional speech databases for the dimensional approach as well [51], [52], [53], [54]. However, in [51], the language is German, in [52], the scope is emotion analysis rather than recognition and in [53], the content consists of nonverbal affect bursts. Finally, we designed our own database that consists of short sections taken from movies. The Semaine database described in [54], is a comprehensive database which includes emotional

1. [18] and [48] contain a review of some commonly used emotional speech databases

ratings of conversations of people with an operator, who adopts in sequence four roles designed to evoke emotional reactions, in many affective dimensions (including arousal and valence). The response format of Semaine database is quite different from that of ANEW. They do not use SAM figures, use a different rating scale that ranges between -1 and 1, and use a trace style continuous rating. If the challenges that the differences in the response format would bring were overcome or ignored, Semaine database would have been the best option within available databases satisfying some of requirements of our work. However, one of the objectives of our study is also to investigate the importance of semantic and acoustic features in arousal and valence dimensions separately. In the database we designed, the clips are rated either just reading the text (subtitles, just semantic features involved), just listening to the audio (acoustic and semantic features involved) or both (acoustic and semantic features involved), which gives us the opportunity that Semaine database lacks, to analyze the effect of semantic and acoustic features on affective dimensions. To our knowledge, there is no available emotional speech database with annotations just using the textual information.

During the selections of the sections from movies , the main difficulty was the existence of some background non-speech audio which is music in most cases. Generally, in movies the intention is to induce feelings that is expressed in the scenes, so a background music that is appropriate for the current feeling is used. We took the sections without this kind of background, that includes monologues or dialogs. We will refer to these sections as 'clips' through this paper. The second difficulty of selecting the clips was to be able to obtain a database including feelings that span all A-V space as much as possible. We tried to have ratings in all four quadrants in A-V space, which are high arousal low valence (HL), high arousal high valence (HH), low arousal low valence (LL) and low arousal high valence (LH) based on the authors' initial judgments (the author rated the clips emotionally to select clips spanning all four quadrants in A-V space, yet, these ratings were not included in the final ratings of each clip).

The response format model is important, since it affects the resultant ratings. The polarity of the format is one example that could affect the ratings. In a format model in which rating 1 and rating 9 represent 'not happy' and 'happy' respectively, if the observers treat 'not happy' as 'sad', it is bipolar. Nevertheless, if the format was aimed for a unipolar rating in which 'not happy' meant 'neutral', the resultant ratings could convey a considerable error due to the confusion. To avoid response format based errors and to coincide with the ANEW work ratings that we use in our results, we used the same response format as in ANEW work. The clips were rated in valence and arousal dimensions in a discrete way using SAM figures. The SAM figures and the instructions given to the participants are provided in [55].

The database consists of 59 clips, the duration of which are between 5 and 25 seconds, from 11 movies in total. The clips include audio and text in the form of subtitles. The audio clips are resampled at 16 kHz. The loudness of the movies may differ a lot recording it. Therefore, a clip with angry, shouting voices and a clip with relaxing, calm voices from two different movies could have similar loudness rates. Since loudness is one of the important acoustics features for emotion speech recognition [18], the long-term loudnesses of the clips have been normalized using Replay Gain[2]. Replay Gain is an open standard loudness calculation algorithm [56] in which the main idea is to calculate the gain needed on an audio file to match the perceived loudness level of a reference audio file.

A Java applet was designed for the experiments to obtain the ratings and was made available online for the participants. There are 3 experiments and a questionnaire in the applet which take around 1 hour in total. In the first experiment, the clips include just text (JT) displayed as subtitles. Once the users finish all 59 clips rating in A-V space, the second experiment is enabled in which there is just audio (JA) parts of the clips. Finally, in the third experiment the clips include both text and audio parts (BTA). The user is provided with the instructions [55], before starting the experiments. The user interface is prepared in a way such that, in the first experiment, the text appears as someone speaks in a clip, but without any sound (subtitle style). After a clip finishes, the user is asked to rate the feeling expressed in the clip, using SAM figures as explained before. The user is not enabled to start the next clip until she/he rates in both dimensions. The user is enabled to listen to a clip more than once that could be useful in cases that he/she could not hear well, or was disturbed, etc. They are also enabled to go back to the previous clip so that they could change their ratings in case they were not sure. Before each experiment, the order of the clips are randomized to make it difficult to remember the ratings for the current clip made in the previous experiments, thus, the order of clips is also different for each participant for each experiment. In each experiment, first two clips are provided just for training purposes. During these two clips they learn how to use the web application, the use of buttons on the application and are able to adjust the volume to their convenience. We are aware of the fact that there could be a learning effect factor affecting the reliability of a rating, since each participant might need different amount of time to get used to the application and the experiment. However, since the order of clips are

2. Please note that the instantaneous loudness in the clips can still be used as an acoustics feature, the normalization is intended just for long-term loudness

different for each experiment and for each participant, this effect is distributed over all the clips, thus, we ignore this effect.

13 people, (7 female, 6 male), between ages 19 and 28, speaking English fluently, were recruited. The participants were able to carry out the experiments from a place they prefer as long as they had Internet with a reliable connection. They were asked to do it in a not disturbing place using a headphone and not to give long breaks between the clips. Their rating times for each clip in each experiment were recorded through the applet. These time recordings were checked for long breaks etc.[3]. In the questionnaire, that the users filled in after finishing the experiments, all of them reported that the instructions were clear and they were confident in rating. The Java code for the applet[4], the start and end times of the clips in each movie (without the audio clips themselves due to copyright restrictions) and the ratings obtained from all participants are provided in [55]. We will call this database as Emotional Movie Database (EMOV) in this paper. Table 1 shows the properties for the EMOV database obtained.

Emotion Model	2 dimensions, Arousal and Valence
Response Format	Rated in A-V space, quantized between 1 and 9
Content	59 short clips from 11 movies in total monologues/dialogues (name of movies and details are in [55])
Duration	5 to 20 seconds
Language	English
Audio	16 kHz, 16 bits
Text	421 words in total (after stop words are extracted)
Annotators	13 people (7 female / 6 male, between ages 19 and 28)

TABLE 1
EMOV Database properties

2.2 Analysis

The reliability of some of the ratings may be low due to various possible experimental errors (e.g. tired or unsure participants). We rejected outliers by applying Peirces outliers detection algorithm [57], which is based on the principle that *the proposed observations should be rejected when the probability of the system of errors obtained by retaining them is less than that of the system of errors obtained by their rejection multiplied by the probability of making so many, and no more, abnormal observations* [58]. 2.9% and 3.7% of data was rejected in total from valence and arousal dimensions respectively (maximum of 2 ratings were excluded for each clip out of 13 ratings and maximum of 7 ratings for excluded for each person out of 59). Figure 1 shows the arithmetic mean of ratings for each clip in A-V space, after rejecting outliers, for all three experiments using JT, JA and BTA. We have ratings in four quadrants (HH,HL,LH,LL), yet there are more data in high arousal low valence quadrant.

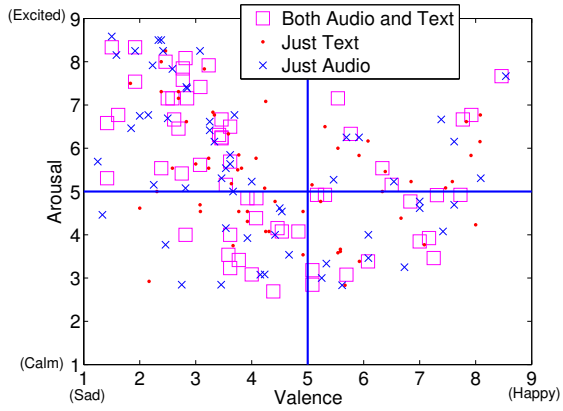


Fig. 1. The arithmetic mean ratings over people of all the clips for the three experiments on A-V space.

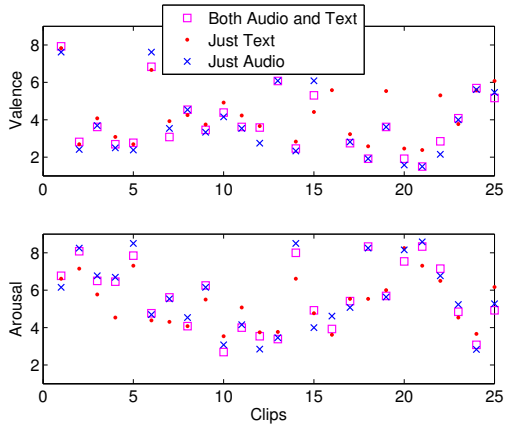


Fig. 2. The arithmetic mean ratings over people of 25 clips for the three experiments for valence and arousal dimensions.

Figure 2 shows the arithmetic mean of ratings for 25 clips (randomly chosen, just for illustration purposes) for valence and arousal dimensions separately, in which differences in ratings for the three experiments can be observed. Figure 3 and 4 show the distributions for all three experiments for 59 clips with arithmetic mean ratings over people for valence and

3. In case of failure to finish all three experiments due to technical difficulties or other reasons, the results have not been included

4. The code can be used to replicate the experiments, or add more data, or make similar experiments with different clips, etc.

arousal dimensions respectively. Table 2 shows the arithmetic mean of the differences in the ratings of the experiments using JT and BTA, and JA and BTA. The results show that the difference between using JA and using BTA is not as high as the difference between using JT and using BTA. In fact, this result is expected since just audio itself already contains the semantics as well since the raters speak English fluently. Meanwhile, the difference in ratings between using JT and BTA gives considerable information on the effect of semantics features on speech emotion analysis. It is observed that the effect differs for valence and arousal dimensions. Figure 5 shows the distribution of the absolute difference between ratings using JT and BTA, for both dimensions. We define a threshold value called 'threshold of acceptable difference (TAD)', that is considered as a boundary between clips with 'high text influence (HTI)' and 'low text influence (LTI)'. We can not determine the value of TAD, however we change it between lowest and highest values possible (zero to three in this case) to analyze the number of clips with HTI and LTI for both dimensions. Figure 6 shows the results of the analysis for both dimensions for HTI clips. It is observed that for valence dimension the number of clips with HTI is considerably higher than that for the arousal dimension. This result implies that using just text part of the speech, people were able to extract the emotion expressed in a clip more easily for the valence dimension than for the arousal dimension.

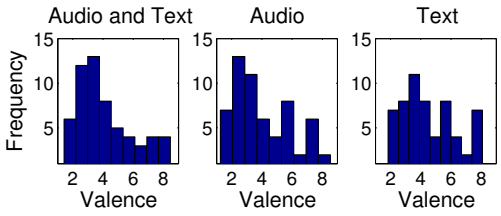


Fig. 3. The distribution of ratings for 59 clips with arithmetic mean ratings over people for the three experiments for the valence dimension.

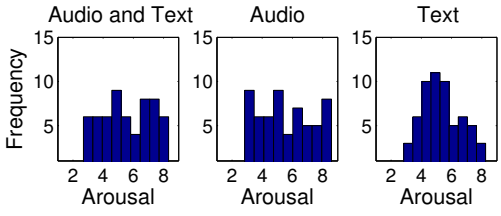


Fig. 4. The distribution of ratings for 59 clips with arithmetic mean ratings over people for the three experiments for the arousal dimension.

	Valence	Arousal
JT-BTA	0.62	0.86
JA-BTA	0.26	0.35

TABLE 2
The arithmetic mean of the differences in the ratings over all clips between the experiments using JT and BTA, and JA and BTA

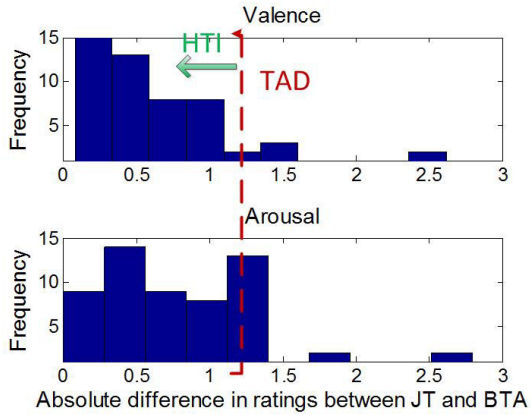


Fig. 5. The distribution of the absolute difference between ratings using JT and BTA (The red line represents TAD)

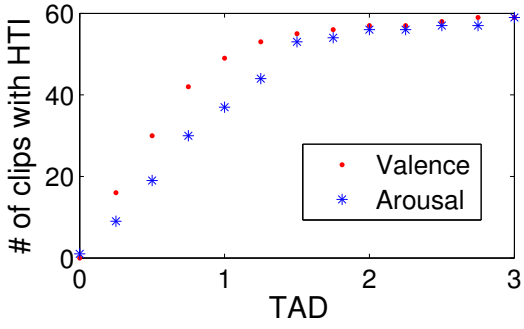


Fig. 6. Number of clips with HTI for arousal and valence dimensions with different TAD values.

3 Modeling Framework

3.1 Overview

The overview of the modeling framework is given in Figure 7. The emotions that the clips convey are recognized using just audio or just text parts separately. In the acoustic part, we extract features, apply a feature selection algorithm and use support vector regression (SVR) on them to estimate A-V values of each clip. In semantic part, we apply latent semantic analysis (LSA) to extract similarities between each word in a clip to an ANEW word. Then we consider

these similarities as weights relating each word to an ANEW word and combine them with A-V values of ANEW words to obtain A-V value of each clip finally. The details of the acoustic and semantic parts will be included in the following sections. Finally, we combine the results of the two parts using Equation 1.

$$AV_{comb}(i) = ws(i) * AV_{sem}(i) + wa(i) * AV_{aco}(i) \quad (1)$$

where, $AV_{comb}(i)$, $AV_{sem}(i)$, $AV_{aco}(i)$ are the results for the combination, semantic and acoustic parts; and $ws(i), wa(i) = (1 - ws(i))$ are the weights of semantic and acoustics parts in the combination, respectively, for the i^{th} clip.

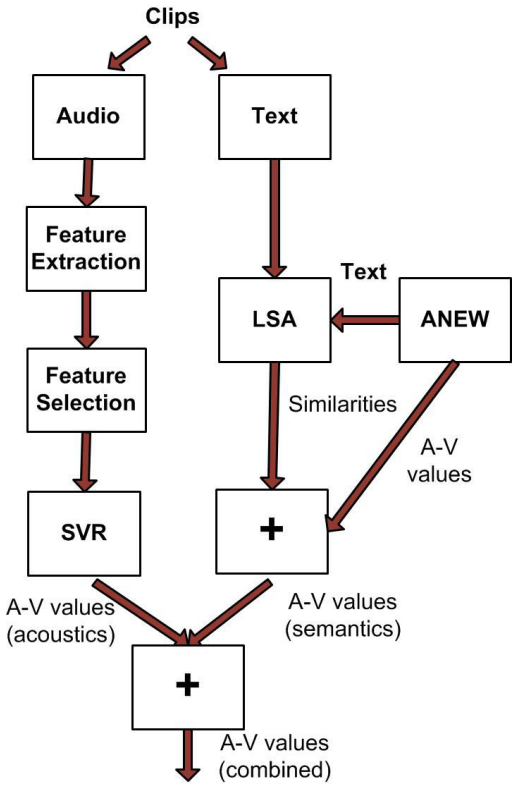


Fig. 7. An overview of the framework. The (+) sign represents that the inputs are combined, the details of the combination methods are in the relative sections.

3.2 Acoustic Analysis

The acoustic part of our work consists of feature extraction, feature selection and regression parts, as stated before, and can be seen in Figure 7. We used the openEar toolkit [7], an open-source affect and emotion recognition engine, that enabled us work in all these three parts using one engine.

For the feature extraction, the openEar toolkit makes use of the SMILE (Speech and Music Interpretation by Large-Space Extraction) [59] that is a signal processing and feature extraction tool and is primarily intended for audio feature extraction. Low level audio features, such as signal energy, pitch, Mel frequency cepstral coefficients (MFCC), perceptual linear predictive coefficients (PLPC) and formants, are extracted and various statistical functionals and transformations, such as extremes, means (arithmetic, geometric, quadratic), moments (variance, skewness etc.), peaks (number of peaks, mean peak distance etc.), are applied to those features. We used 988 features in total.

As a feature selection method we use the correlation feature selection (CFS) which is based on the hypothesis that good features are the ones which do not correlate with each other but correlate with the classification. Finally, as the recognition engine we have SVR which is a maximum margin algorithm, computing a linear regression function in a high dimensional kernel induced feature space where the input data is using a nonlinear transformation [6]. The details about CFS and SVR can be found in openEar toolkit [7].

3.3 Semantic Analysis

1034 ANEW words were rated by humans on three affective dimensions, valence, arousal and dominance. We use the ratings of these words on 2 dimensions, valence and arousal, to map the words in the clips on A-V space. In other words, ANEW words are used as emotional keywords as mentioned earlier. We do not look for these exact keywords in the text of the clips, but estimate the A-V values of each word in a clip using the similarities of them to the ANEW words. We apply latent semantic analysis (LSA) to calculate these similarities. LSA is an indexing method that is based on the principle that words occurring in similar contexts are also similar in meanings [8]. An LSA software package [60], that is based on term frequency inverse document frequency (TF-IDF) weighting and outputs cosine distance as the similarity measure, has been used. The corpus used for LSA is called HAWIK combining Harvard Classics literature samples with Wikipedia articles and Reuters news items. HAWIK corpus has been used in [40] and has been shown to perform well for affect recognition purposes. The database we obtained has been filtered out of 'stop-words' which are used often but have no effect semantically like 'the', 'a', 'are', etc. The 'stop-word' list we used is provided in [55]. The following formula can be used to find the estimated A-V values of the corresponding words,

$$AV_{word}(j) = \frac{\sum_{i=1}^{N} AV_{ANEW}(ji) * sim_{ANEW}(ji)}{\sum_{i=1}^{N} sim_{ANEW}(ji)}, \quad (2)$$

where $AV_{ANEW}(ji)$, $sim_{ANEW}(ji)$ are the A-V values of the ANEW words and the cosine similarities of the corresponding word (as a result of the LSA algorithm) to ANEW words for the j^{th} clip word to be analyzed and i^{th} ANEW word respectively. N is the number of ANEW words giving similarities greater than a threshold value of sim_{thr}, that is to be optimized. Thus, N ANEW words with similarities to clip words above the threshold are taken into account within this formula. However, there is a trade-off choosing this threshold value. If we choose it to be low, in the case of a word having a similarity of 1 to one of the ANEW words, we would be estimating the A-V value of it using a high number of words (large N), which could lead to redundancy. If we choose it to be high, in the case of a clip word with relatively low similarities to all ANEW words, it might result in N to be zero, so we would not be able to estimate the A-V value of the word. Thus, instead of using the cosine similarity, calculated by applying LSA, directly; we define a weight for each word using Equation 3, which solves this problem by adjusting the threshold, sim_{thr}, using the information of the maximum similarity of a word to any ANEW word,

$$we_{ANEW}(ji) = sim_{ANEW}(ji)^{\frac{th}{1+gamma-sim_{max}(j)}} \quad (3)$$

where $we_{ANEW}(ji)$, th and $sim_{max}(j)$ are the weight for each ANEW word for the j^{th} clip word, a threshold value to be optimized between 0 and 1 and the maximum similarity of j^{th} clip word to ANEW words respectively. $gamma$ is a very small number (close to zero) used to avoid dividing by zero in the case of sim_{max} of 1. Figure 8 illustrates the weight-similarity relation visually. The red line in the figure, which gives weight of zero to all similarities below zero, shows the case when the maximum cosine similarity of a word to ANEW words is 1. In this case, only the A-V values of the corresponding ANEW word is taken as desired. Then, we insert $we_{ANEW}(ji)$ instead of $sim_{ANEW}(ji)$ in Equation 2 to estimate AV values of a word in a clip. Finally, to estimate the A-V value of a clip, we use the estimated A-V value of each word and calculate the weighted average, taking the maximum similarity score found for each word as its weight.

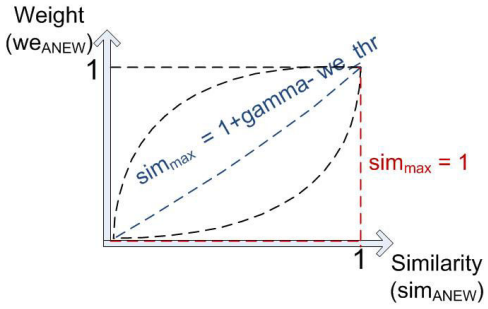


Fig. 8. Similarity to weight conversion using Equation 3. The curve shape changes with th and sim_{max}. The red and blue colored curves show two different specific cases.

4 Experiments and Results

4.1 Acoustic Experiments

For acoustic features, the frame size is set to 25 ms at a skip rate of 10 ms. The frequency range of the spectrum is set from 0 to 8 kHz. We used 988 features in total including signal energy, pitch, MFCC, PLPC and formants and their variants with statistical functionals and transformations applied on, as stated in section 3.2.

We apply correlation feature selection (CFS) on all acoustic features obtained for both dimensions. For valence dimension around 40 features remain after applying CFS, while for the arousal dimension the number of features remain is 70. This difference mainly results from the existence of additional loudness related features for the arousal dimension. We use a radial basis function (RBF) type kernel for the valence dimension while the linear type kernel for the arousal dimension, since the number of data samples is lower than features for the arousal dimension. The details about all the parameters for the acoustics part can be found in [7].

All 59 clips have been used to find the optimal parameters using 5-fold cross validation method of the openEar toolkit. The final results are given using the leave-one-out method. We measure the error between the estimated and human-rated A-V values using mean absolute error (MAE) which gives the average of the absolute differences between them. We also check the root mean squared error (RMSE) for the final results which is a useful error measure for the applications where large errors might be specifically undesirable.[5] Finally, we check for normalized mean absolute error (NMAE) and normalized root mean squared error (NMRSE) which are simply calculated by dividing MAE and RMSE by the difference of the maximum and minimum data (arousal and valence ratings) values observed [6]. Table 3 gives the resultant errors (the database has been divided into a train set with 29 clips and a test set with 30 clips and the results shown in this table are for the test set for comparison purposes).

4.2 Semantic Experiments

The database has been divided into a train set with 29 clips and a test set with 30 clips. Optimization of the

5. The errors are squared before they are averaged, hence, higher weight is given for large differences.

6. $NMAE/NRSME = \frac{MAE/NRMSE}{(AV_{max}-AV_{min})}$, where AV_{max} and AV_{min} represent the maximum and minimum AV ratings in the database

	MAE	RMSE	NMAE	NRMSE
Valence	1.98	2.50	0.28	0.36
Arousal	1.29	1.54	0.23	0.27

TABLE 3
Mean Absolute error (MAE), Root Mean Squared Error (RMSE), Normalized Mean Absolute error (NMAE) and Normalized Root Mean Squared Error (NRMSE) for acoustic part in A-V dimensions, using the test set

	MAE	RMSE	NMAE	NRMSE
Valence	1.45	1.85	0.21	0.26
Arousal	1.39	1.64	0.25	0.29

TABLE 4
Mean Absolute error (MAE), Root Mean Squared Error (RMSE), Normalized Mean Absolute error (NMAE) and Normalized Root Mean Squared Error (NRMSE) for semantic part in A-V dimensions

parameters are done using the train set and the final evaluation is done using the test set. As in the acoustic part, we use MAE and RMSE as error measures. The optimal values found for sim_thr and th, in equations 2 and 3, are 0.2 and 0.3 respectively.

We looked for a subset of ANEW words that could perform better at recognizing affect. We modeled our framework in such a way that ANEW words, which were already rated in A-V space by humans, could be the reference part semantically. Nevertheless, we did not know if the choice of those specific words are the optimum for our work. Therefore, we created a subset of ANEW words (116 words the index of which could be found in [55]), which are affect related like adjectives 'afraid', 'depressed' or nouns like 'fear' and 'hate'. Figure 9 gives the MAE results (RMSE results are similar) of the train set using the subset and all ANEW words using the train set. We use the subset to obtain the final results since their performance surpasses all ANEW words' performance for our work. Table 4 gives the results for the semantics part using the test set with optimal parameter values.

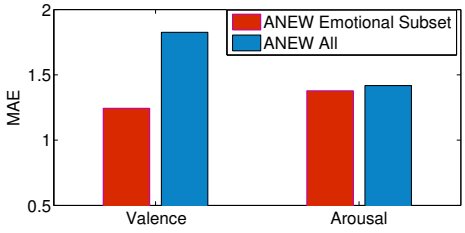


Fig. 9. Mean Absolute error (MAE) results for both dimensions using whole ANEW words or the emotional words subset, using the train set.

4.3 Combination of Acoustic and Semantic Experiments

We searched the optimal weights for semantic and acoustic parts (ws and wa as in Equation 2) for the combination of the results of the two parts, giving minimum MAE and RMSE using the train set. Figure 10 gives the results using ws values between 0 and 1. Thus, we choose ws, giving minimum MAE and RMSE, to be 0.8 and 0.85 for the valence dimension, and 0 and 0.2 for the arousal dimension respectively (then, wa is simply equal to (1-ws)).

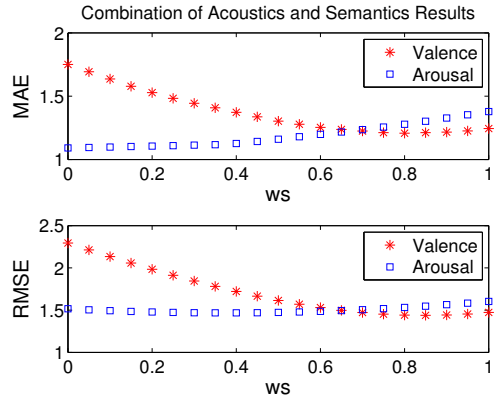


Fig. 10. Mean Absolute error (MAE) and Root Mean Squared Error (RMSE) versus the weight of semantics (ws) in the combination, using the train set.

We are mainly interested in the results obtained by combining them to see if we have the desirable and expected improvement and the effect of it in the two dimensions. For the weights of semantics and acoustics in A-V dimensions, we use the optimal values found as described in the previous paragraph and as shown in Figure 10. Table 5 shows the results using the test set. Figure 11 shows the results for each clip in test set on A-V space. Figure 12 shows the arithmetic mean of ratings for 15 clips from the test set for valence and arousal dimensions separately, in which differences in ratings for different cases can be observed.

5 Discussion

Since the dimensional affect recognition research area is quite recent and since we created our own database, it is arduous to make a comparison between our results and the results obtained in previous works. Nevertheless, the reported results for the SAL corpus, that also has ratings of speech data on A-V space

		Weights (ws / wa)	Combined Result
Valence	MAE	0.80 / 0.20	1.40
	RMSE	0.85 / 0.15	1.77
	NMAE	0.80 / 0.20	0.20
	NRMSE	0.85 / 0.15	0.25
Arousal	MAE	0 / 1	1.28
	RMSE	0.20 / 0.80	1.52
	NMAE	0 / 1	0.23
	NRMSE	0.20 / 0.80	0.26

TABLE 5
Mean Absolute error (MAE), Root Mean Squared Error (RMSE), Normalized Mean Absolute error (NMAE) and Normalized Root Mean Squared Erro (NRMSE) for the combination of semantic and acoustic parts in A-V dimensions with weight values each part, using the test set

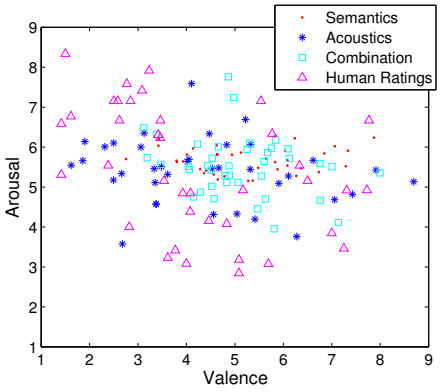


Fig. 11. The human ratings and regression results using just semantic features, just acoustic features or the combination of them, in A-V space.

with a rating scale between -1 and 1, in [7], are 0.28 and 0.38 of MAE for arousal and valence dimensions respectively. If we normalize their results to our rating scale that is between 1 and 9, the results would be 1.12 and 1.52 for A-V dimensions which is comparable to our final results as seen in Table 5 that are 1.28 and 1.40 respectively.

In this study, we investigated the effect of combining semantics and acoustics features for affect recognition from speech, if and how much the performance increases after the combination, in a dimensional perspective. We observe, as seen in Table 5, that the results are slightly better in both dimensions than the best of semantic and acoustic results (valence dimension result from 1.45 using just semantic and 1.98 using just acoustic to 1.40 using both semantic and acoustic features, and arousal dimension result from 1.39 using just semantic and 1.29 using just acoustic to 1.29 using both semantic and acoustic features).

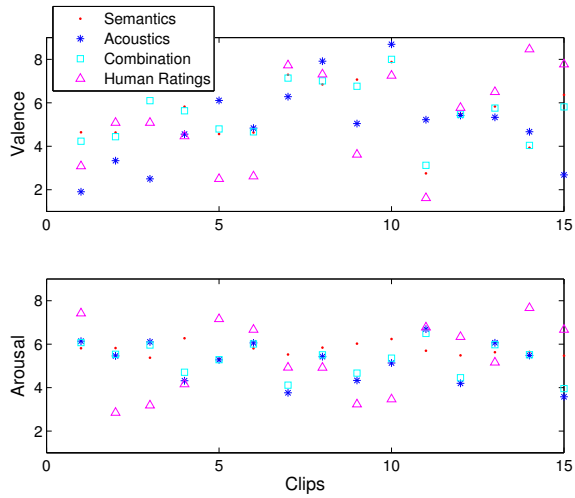


Fig. 12. The arithmetic mean ratings over people of 15 clips from the test set for the three experiments for valence and arousal dimensions.

Therefore, our results show that combining acoustic and semantic information improves the recognition results for dimensional approach as it was shown before for both the categorical and the dimensional approach in previous works [37], [20].

The most interesting and a novel result in our work is that we observe and show that the importance of acoustic and semantic features differ for arousal and valence dimensions. The weight of the semantic features is much higher for valence dimension while the weight of acoustic features is much higher for arousal dimension as seen in Table 5. We show that the valence dimension is recognized better using semantic features while the arousal dimension using acoustics features. These results also coincides with the human subjects' ratings for the database, explained in section 2, from which we inferred that it was easier for humans to extract the emotion expressed in a clip in the valence dimension using just text part of the speech. We interpret the result as, *the valence dimension is more about what we say, while the arousal dimension is more about how we say it*. This conclusion is also coherent with the fact that it has not been shown yet how or if the acoustics features correlate with the valence dimension [18] and also with results of previous studies on emotion recognition ([10], [61]). In [10], it is claimed that vocal characteristics are sensitive to only arousal dimension. In [61], where they design an emotion recognition model in A-V space as well using bio-sensors extracting features, such as body temperature, breath speed and heart activation, they show that it is much harder to estimate the valence dimension than the arousal. In other words, they show that physiological changes in

the body gives more information about the arousal dimension. Therefore, since, the physiological changes affect our speech as well, specifically how we speak (higher arousal induces greater sub-glottal pressure resulting in change in speech as well such as making it louder, faster and have higher average pitch, etc.), they support our result that acoustics features are more important for the arousal dimension.

6 Conclusion

This paper investigated the dimensional emotion recognition in speech using both audio and textual information. The aim of the study was also to investigate the effect of semantic and acoustic features for arousal and valence dimensions. We designed our own emotional speech database using short movie clips of 5 to 25 seconds of duration, rated by human subjects in A-V space. Firstly, we recognized emotions using acoustic and semantic features separately. We combined many acoustics features including MFCC, PLPC, pitch, formants, etc., applied correlation based feature selection and support vector regression to estimate A-V values of each speech sample. For semantic part, we used affective norms for English words (ANEW), including 1034 English words rated by human subjects in A-V space, as keywords, and applied latent semantics analysis (LSA) to calculate the cosine similarities between those ANEW words and words in the database, and used the results to estimate the A-V values of each clip. Finally, we combined the results of acoustic and semantic parts, assigning different weights to two parts, for each dimension. We showed that the integration of semantic and acoustic features improve the performance of dimensional speech emotion recognition and the weight of each part differed for valence and arousal dimensions. We concluded that *the valence dimension is more about what we say, while the arousal dimension is more about how we say it*.

Acknowledgments

This work was supported in part by the IST Programme of the European Community, under the PASCAL2 Network of Excellence, IST-2007-216886 and in part by the Danish Council for Strategic Research of the Danish Agency for Science Technology and Innovation under the CoSound project, case number 11-115328. The publication only reflects the authors' views.

References

[1] R. Picard, E. Vyzas, and J. Healey, "Toward machine emotional intelligence: Analysis of affective physiological state," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 10, pp. 1175–1191, 2001.

[2] S. Karadoğan and J. Larsen, "Combining semantics and acoustic features for valence and arousal recognition of speech," in *The Third International Workshop on Cognitive Information Processing*, 2012.

[3] M. Bradley and P. Lang, "Affective norms for english words (anew): Instruction manual and affective ratings," *University of Florida: The Center for Research in Psychophysiology*, 1999.

[4] Bradley, M.M. and Lang, P.J., "Measuring emotion: the self-assessment manikin and the semantic differential," *Journal of behavior therapy and experimental psychiatry*, vol. 25, no. 1, pp. 49–59, 1994.

[5] M. Hall, "Correlation-based feature selection for machine learning," Ph.D. dissertation, The University of Waikato, 1999.

[6] A. Smola and B. Schölkopf, "A tutorial on support vector regression," *Statistics and computing*, vol. 14, no. 3, pp. 199–222, 2004.

[7] F. Eyben, M. Wollmer, and B. Schuller, "Openear introducing the munich open-source emotion and affect recognition toolkit," in *3rd International Conference on Affective Computing and Intelligent Interaction and Workshops*. IEEE, 2009, pp. 1–6.

[8] T. Landauer, P. Foltz, and D. Laham, "An introduction to latent semantic analysis," *Discourse processes*, vol. 25, no. 2-3, pp. 259–284, 1998.

[9] C. Darwin, *The expression of the emotions in man and animals*. University of Chicago Press (original work published in 1872), 1965.

[10] I. Mauss and M. Robinson, "Measures of emotion: A review," *Cognition and Emotion*, vol. 23, no. 2, pp. 209–237, 2009.

[11] H. Gunes and M. Pantic, "Automatic, dimensional and continuous emotion recognition," *International Journal of Synthetic Emotions*, vol. 1, no. 1, pp. 68–99, 2010.

[12] P. Ekman and W. Friesen, "Constants across cultures in the face and emotion." *Journal of personality and social psychology*, vol. 17, no. 2, p. 124, 1971.

[13] P. Ekman, "An argument for basic emotions," *Cognition & Emotion*, vol. 6, no. 3-4, pp. 169–200, 1992.

[14] P. Lang, M. Bradley, B. Cuthbert *et al.*, "Motivated attention: Affect, activation, and action," *Attention and orienting: Sensory and motivational processes*, pp. 97–135, 1997.

[15] L. Barrett and J. Russell, "The structure of current affect," *Current Directions in Psychological Science*, vol. 8, no. 1, p. 10, 1999.

[16] J. Russell and A. Mehrabian, "Evidence for a three-factor theory of emotions," *Journal of research in Personality*, vol. 11, no. 3, pp. 273–294, 1977.

[17] R. Fernandez, "A computational model for the automatic recognition of affect in speech," Ph.D. dissertation, Massachusetts Institute of Technology, 2004.

[18] M. El Ayadi, M. Kamel, and F. Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," *Pattern Recognition*, vol. 44, no. 3, pp. 572–587, 2011.

[19] S. Zhang, Q. Tian, S. Jiang, Q. Huang, and W. Gao, "Affective mtv analysis based on arousal and valence features," in *IEEE International Conference on Multimedia and Expo*. IEEE, 2008, pp. 1369–1372.

[20] B. Schuller, "Recognizing affect from linguistic information in 3d continuous space," *IEEE Transactions on Affective Computing*, no. 99, pp. 1–1, 2011.

[21] M. Faith and J. Thayer, "A dynamical systems interpretation of a dimensional model of emotion," *Scandinavian Journal of Psychology*, vol. 42, no. 2, pp. 121–133, 2001.

[22] T. Nwe, S. Foo, and L. De Silva, "Speech emotion recognition using hidden markov models," *Speech communication*, vol. 41, no. 4, pp. 603–623, 2003.

[23] D. Ververidis and C. Kotropoulos, "Emotional speech classification using gaussian mixture models and the sequential floating forward selection algorithm," in *IEEE International Conference on Multimedia and Expo*. IEEE, 2005, pp. 1500–1503.

[24] Ververidis, D. and Kotropoulos, C., "Emotional speech recognition, resources, features, and methods," *Speech communication*, vol. 48, no. 9, pp. 1162–1181, 2006.

[25] B. Schuller, S. Reiter, R. Muller, M. Al-Hames, M. Lang, and G. Rigoll, "Speaker independent speech emotion recognition by ensemble classification," in *IEEE International Conference on Multimedia and Expo*. IEEE, 2005, pp. 864–867.

[26] C. Lee and S. Narayanan, "Toward detecting emotions in spoken dialogs," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 2, pp. 293–303, 2005.
[27] A. Batliner, S. Steidl, B. Schuller, D. Seppi, T. Vogt, J. Wagner, L. Devillers, L. Vidrascu, V. Aharonson, L. Kessous *et al.*, "Whodunnit-searching for the most important feature types signalling emotion-related user states in speech," *Computer Speech & Language*, vol. 25, no. 1, pp. 4–28, 2011.
[28] O. Pierre-Yves, "The production and recognition of emotions in speech: features and algorithms," *International Journal of Human-Computer Studies*, vol. 59, no. 1, pp. 157–183, 2003.
[29] T. Vogt and E. André, "Comparing feature sets for acted and spontaneous speech in view of automatic emotion recognition," in *IEEE International Conference on Multimedia and Expo*. IEEE, 2005, pp. 474–477.
[30] A. Prasad, Y. Srinivas, P. Brahmaiah *et al.*, "Gender based emotion recognition system for telugu rural dialects using hidden markov models," *Arxiv preprint arXiv:1006.4548*, 2010.
[31] C. Yu, Q. Tian, F. Cheng, and S. Zhang, "Speech emotion recognition using support vector machines," *Advanced Research on Computer Science and Information Engineering*, pp. 215–220, 2011.
[32] B. Han, S. Ho, R. Dannenberg, and E. Hwang, "Smers: Music emotion recognition using support vector regression," in *Proceedings of the 10th International Conference on Music Information Retrieval (ISMIR 2009)*, 2009.
[33] E. Kim, K. Hyun, S. Kim, and Y. Kwak, "Improved emotion recognition with a novel speaker-independent feature," *IEEE/ASME Transactions on Mechatronics*, vol. 14, no. 3, pp. 317–325, 2009.
[34] F. Eyben, M. Wöllmer, A. Graves, B. Schuller, E. Douglas-Cowie, and R. Cowie, "On-line emotion recognition in a 3-d activation-valence-time continuum using acoustic and linguistic cues," *Journal on Multimodal User Interfaces*, vol. 3, no. 1, pp. 7–19, 2010.
[35] S. Zhang, X. Zhao, and B. Lei, "Spoken emotion recognition using radial basis function neural network," *Advances in Computer Science, Environment, Ecoinformatics, and Education*, pp. 437–442, 2011.
[36] B. Schuller, G. Rigoll, and M. Lang, "Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine-belief network architecture," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1. IEEE, 2004, pp. I–577.
[37] Z. Chuang and C. Wu, "Multi-modal emotion recognition from speech and text," *International Journal of Computational Linguistics and Chinese Language Processing*, vol. 9, no. 2, pp. 45–62, 2004.
[38] A. Batliner, S. Steidl, B. Schuller, D. Seppi, K. Laskowski, T. Vogt, L. Devillers, L. Vidrascu, N. Amir, L. Kessous *et al.*, "Combining efforts for improving automatic classification of emotional user states," *Proc. IS-LTC*, pp. 240–245, 2006.
[39] C. Wu, Z. Chuang, and Y. Lin, "Emotion recognition from text using semantic labels and separable mixture models," *ACM Transactions on Asian Language Information Processing (TALIP)*, vol. 5, no. 2, pp. 165–183, 2006.
[40] M. Petersen and L. Hansen, "Modeling lyrics as emotional semantics," *Proceedings of YoungCT, KAIST Korea Advanced Institute of Science and Technology*, 2010.
[41] A. Gill, R. French, D. Gergle, and J. Oberlander, "The language of emotion in short blog texts," in *Proceedings of the 2008 ACM conference on Computer supported cooperative work*. ACM, 2008, pp. 299–302.
[42] C. Strapparava and R. Mihalcea, "Learning to identify emotions in text," in *Proceedings of the 2008 ACM symposium on Applied computing*. ACM, 2008, pp. 1556–1560.
[43] S. Yildirim, S. Narayanan, and A. Potamianos, "Detecting emotional state of a child in a conversational computer game," *Computer Speech & Language*, vol. 25, no. 1, pp. 29–44, 2011.
[44] R. Stevenson, J. Mikels, and T. James, "Characterization of the affective norms for english words by discrete emotional categories," *Behavior research methods*, vol. 39, no. 4, pp. 1020–1024, 2007.
[45] P. Lewis, H. Critchley, P. Rotshtein, and R. Dolan, "Neural correlates of processing valence and arousal in affective words," *Cerebral Cortex*, vol. 17, no. 3, p. 742, 2007.
[46] J. Redondo, I. Fraga, I. Padrón, and M. Comesaña, "The spanish adaptation of anew (affective norms for english words)," *Behavior research methods*, vol. 39, no. 3, pp. 600–605, 2007.
[47] Y. Lin and G. Wei, "Speech emotion recognition based on hmm and svm," in *Proceedings of International Conference on Machine Learning and Cybernetics*, vol. 8. IEEE, 2005, pp. 4898–4901.
[48] E. Douglas-Cowie, N. Campbell, R. Cowie, and P. Roach, "Emotional speech: Towards a new generation of databases," *Speech Communication*, vol. 40, no. 1-2, pp. 33–60, 2003.
[49] University of Pennsylvania Linguistic Data Consortium, "Emotional prosody speech and transcripts," 2002, `http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2002S28S`.
[50] F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, and B. Weiss, "A database of german emotional speech," in *Ninth European Conference on Speech Communication and Technology*, 2005.
[51] M. Grimm, K. Kroschel, and S. Narayanan, "The vera am mittag german audio-visual emotional speech database," in *IEEE International Conference on Multimedia and Expo*. IEEE, 2008, pp. 865–868.
[52] E. Douglas-Cowie, R. Cowie, I. Sneddon, C. Cox, O. Lowry, M. Mcrorie, J. Martin, L. Devillers, S. Abrilian, A. Batliner *et al.*, "The humaine database: addressing the collection and annotation of naturalistic and induced emotional data," *Affective computing and intelligent interaction*, pp. 488–500, 2007.
[53] P. Belin, S. Fillion-Bilodeau, and F. Gosselin, "The montreal affective voices: A validated set of nonverbal affect bursts for research on auditory affective processing," *Behavior research methods*, vol. 40, no. 2, pp. 531–539, 2008.
[54] G. McKeown, M. Valstar, R. Cowie, and M. Pantic, "The semaine corpus of emotionally coloured character interactions," in *IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2010, pp. 1079–1084.
[55] Karadogan, S. G., "Emotional speech database design details and the java applet code written for the design," 2012, www.imm.dtu.dk/pubdb/p.php?6231.
[56] D. Robinson, "Replay gain," 2001, http://www.replaygain.org/.
[57] S. Ross, "Peirce's criterion for the elimination of suspect experimental data," *Journal of Engineering Technology*, vol. 20, no. 2, pp. 38–41, 2003.
[58] B. Peirce, "Criterion for the rejection of doubtful observations," *The Astronomical Journal*, vol. 2, pp. 161–163, 1852.
[59] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," in *Proceedings of the International Conference on Multimedia*. ACM, 2010, pp. 1459–1462.
[60] DTU, "Java LSA Software with HAWIK Corpus Matrices," 2010, www.imm.dtu.dk/pubdb/p.php?6320.
[61] A. Haag, S. Goronzy, P. Schaich, and J. Williams, "Emotion recognition using bio-sensors: First steps towards an automatic system," *Affective Dialogue Systems*, pp. 36–48, 2004.

Seliz G. Karadogan Seliz Karadogan received the BSc degree from Middle East Technical University (METU) in 2006, and the MSc degree from École Polytechnique Fédérale de Lausanne (EPFL) in 2009. She is a PhD student at the Department of Informatics, Technical University of Denmark (DTU), supervised by Assoc. Prof. Jan Larsen. Her PhD project has a title of 'Cognizant Hearing Aids' and she is working on and interested in audio processing applications, speech recognition, affective computing and machine learning for signal processing applications. She is a member of The Institute of Electrical and Electronics Engineers (IEEE), Pattern Analysis, Statistical Modeling and Computational Learning (PASCAL) and The HUMAINE Association.

Jan Larsen Jan Larsen received the M.Sc. and Ph.D. degrees in electrical engineering from the Technical University of Denmark (DTU) in 1989 and 1994. Dr. Larsen is Associate Professor of Digital Signal Processing at Department of Informatics, DTU. Jan Larsen has authored and co-authored more than 125 papers and book chapters within the areas of nonlinear statistical signal processing, machine learning, neural networks and datamining with applications to biomedicine, monitoring systems, multimedia, audio, and webmining. According to Google Scholar the no. of citations is 2096 and the h-index is 26 as of 27.02.2012. He has participated in more than ten national and international research programs, and has served as reviewer for many international journals, conferences, publishing companies and research funding organizations. As regards synergistic activities he took part in conference organizations, among others the IEEE Workshop on Machine Learning for Signal Processing (formerly Neural Networks for Signal Processing) 1999-2012. He is Director of Network for Danish Sound Technology (2009-2014), past chair of the IEEE Machine Learning for Signal Processing Technical Committee of the IEEE Signal Processing Society (2005-2007), and chair of IEEE Denmark Section's Signal Processing Chapter (2002-). He is a senior member of The Institute of Electrical and Electronics Engineers. Other professional committee participation includes: Steering committee member of the Audio Signal Processing Network in Denmark, 2006-. Editorial Board Member of Signal Processing, Elsevier, 2006-2007; and guest editorships involves IEEE Transactions on Neural Networks; Journal of VLSI Signal Processing Systems; and Neurocomputing.

Bibliography

[AAZ+10] M.A.M. Abushariah, R.N. Ainon, R. Zainuddin, M. Elshafei, and O.O. Khalifa. Natural speaker-independent arabic speech recognition system based on hidden markov models using sphinx tools. In *Computer and Communication Engineering (ICCCE), 2010 International Conference on*, pages 1–6. IEEE, 2010.

[ACMR00] C.M. Arrington, T.H. Carr, A.R. Mayer, and S.M. Rao. Neural mechanisms of visual attention: object-based selection of a region in space. *Journal of Cognitive Neuroscience*, 12(Supplement 2):106–117, 2000.

[ACS09] O. AlZoubi, R. Calvo, and R. Stevens. Classification of eeg for affect recognition: An adaptive approach. *AI 2009: Advances in Artificial Intelligence*, pages 52–61, 2009.

[Ake08] M.A. Akeroyd. Are individual differences in speech reception related to individual differences in cognitive ability? a survey of twenty experimental studies with normal and hearing-impaired adults. *International journal of audiology*, 47(S2):53–71, 2008.

[ALLKPF09] S. Arlinger, T. Lunner, B. Lyxell, and M. KATHLEEN PICHORA-FULLER. The emergence of cogni-

tive hearing science. *Scandinavian journal of psychology*, 50(5):371–384, 2009.

[BAS09] T. Baumann, M. Atterer, and D. Schlangen. Assessing and improving the performance of speech recognition for incremental systems. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 380–388. Association for Computational Linguistics, 2009.

[BCS+99] M. Beutnagel, A. Conkie, J. Schroeter, Y. Stylianou, and A. Syrdal. The at&t next-gen tts system. In *Joint Meeting of ASA, EAA, and DAGA*, pages 18–24. Citeseer, 1999.

[BCTW12] J. Benichov, L.C. Cox, P.A. Tun, and A. Wingfield. Word recognition within a linguistic context: Effects of age, hearing acuity, verbal ability, and cognitive function. *Ear and Hearing*, 33(2):250, 2012.

[Bec88] W. Bechtel. *Philosophy of science: An overview for cognitive science*. Lawrence Erlbaum, 1988.

[BFBG08] P. Belin, S. Fillion-Bilodeau, and F. Gosselin. The montreal affective voices: A validated set of nonverbal affect bursts for research on auditory affective processing. *Behavior research methods*, 40(2):531–539, 2008.

[BGH10] A. Birlutiu, P. Groot, and T. Heskes. Multi-task preference learning with an application to hearing aid personalization. *Neurocomputing*, 73(7-9):1177–1185, 2010.

[BHM10] A. Bender, E. Hutchins, and D. Medin. Anthropology in cognitive science. *Topics in Cognitive Science*, 2(3):374–385, 2010.

[BL94] M.M. Bradley and P.J. Lang. Measuring emotion: the self-assessment manikin and the semantic differential. *Journal of behavior therapy and experimental psychiatry*, 25(1):49–59, 1994.

[BL99] M.M. Bradley and P.J. Lang. Affective norms for english words (anew): Instruction manual and affective ratings. *University of Florida: The Center for Research in Psychophysiology*, 1999.

[BM09] A. Bertrand and M. Moonen. Robust distributed noise reduction in hearing aids with external acoustic sensor nodes. *EURASIP Journal on Advances in Signal Processing*, 2009:12, 2009.

[Bod06] M.A. Boden. *Mind as machine: a history of cognitive science*, volume 1. Clarendon press, 2006.

[Bol79] S. Boll. Suppression of acoustic noise in speech using spectral subtraction. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 27(2):113–120, 1979.

[BPR+05] F. Burkhardt, A. Paeschke, M. Rolfes, W.F. Sendlmeier, and B. Weiss. A database of german emotional speech. In *Ninth European Conference on Speech Communication and Technology*, 2005.

[BR99] L.F. Barrett and J.A. Russell. The structure of current affect. *Current Directions in Psychological Science*, 8(1):10, 1999.

[Bre94a] A.S. Bregman. *Auditory scene analysis: The perceptual organization of sound*. The MIT Press, 1994.

[Bre94b] A.S. Bregman. *Auditory scene analysis: The perceptual organization of sound*. The MIT Press, 1994.

[Bro58] D. E. Broadbent. *Perception and Communication*. Pergamon Press, 1958.

[Bro00] A.W. Bronkhorst. The cocktail party phenomenon: A review of research on speech intelligibility in multiple-talker conditions. *Acta Acustica united with Acustica*, 86(1):117–128, 2000.

[BSS+06] A. Batliner, S. Steidl, B. Schuller, D. Seppi, K. Laskowski, T. Vogt, L. Devillers, L. Vidrascu,

N. Amir, L. Kessous, et al. Combining efforts for improving automatic classification of emotional user states. *Proc. IS-LTC*, pages 240–245, 2006.

[BSS+11] A. Batliner, S. Steidl, B. Schuller, D. Seppi, T. Vogt, J. Wagner, L. Devillers, L. Vidrascu, V. Aharonson, L. Kessous, et al. Whodunnit-searching for the most important feature types signalling emotion-related user states in speech. *Computer Speech & Language*, 25(1):4–28, 2011.

[BSW89] A.G. Barto, R.S. Sutton, and C.J.C.H. Watkins. Learning and sequential decision making. In *Learning and computational neuroscience*. Citeseer, 1989.

[BZM98] M. Behrmann, R.S. Zemel, and M.C. Mozer. Object-based attention and occlusion: evidence from normal participants and a computational model. *Journal of Experimental Psychology: Human Perception and Performance*, 24(4):1011, 1998.

[CCB01] A.R.A. Conway, N. Cowan, and M.F. Bunting. The cocktail party phenomenon revisited: The importance of working memory capacity. *Psychonomic Bulletin & Review*, 8(2):331–335, 2001.

[Che53] E. C. Cherry. Some experiments on the recognition of speech, with one and two ears. *Journal pf Acoustic Society of America*, 25:975–979, 1953.

[Che12] C.C. Cheng. *Online Learning of Large Margin Hidden Markov Models for Automatic Speech Recognition*. PhD thesis, University of California, San Diego, 2012.

[Cow01] N. Cowan. The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and brain sciences*, 24(01):87–114, 2001.

[CW04] Z.J. Chuang and C.H. Wu. Multi-modal emotion recognition from speech and text. *International Journal of Computational Linguistics and Chinese Language Processing*, 9(2):45–62, 2004.

[DA07] V. Duangudom and D.V. Anderson. Using auditory saliency to interpret complex auditory scenes. *Journal of the Acoustical Society of America*, 121(5):3119, 2007.

[DAM11] A.O. Diaconescu, C. Alain, and A.R. McIntosh. Modality-dependent 'what' and 'where' preparatory processes in auditory and visual systems. *Journal of Cognitive Neuroscience*, 23(7):1609–1623, 2011.

[Dar65] C. Darwin. 1872 the expression of the emotions in man and animals, 1965.

[DASW94] L. Deng, M. Aksmanovic, X. Sun, and C.F.J. Wu. Speech recognition using hidden markov models with polynomial regression functions as nonstationary states. *IEEE Transactions on Speech and Audio Processing*, 2(4):507–520, 1994.

[Dau09] T. Dau. Auditory processing models. *Handbook of Signal Processing in Acoustics*, pages 175–196, 2009.

[DCB08] B. De Coensel and D. Botteldooren. Modeling auditory attention focusing in multisource environments. In *Proceedings of the Acoustics' 08 Conference*, page 3255, 2008.

[DCCCR03] E. Douglas-Cowie, N. Campbell, R. Cowie, and P. Roach. Emotional speech: Towards a new generation of databases. *Speech Communication*, 40(1-2):33–60, 2003.

[DCCS+07] E. Douglas-Cowie, R. Cowie, I. Sneddon, C. Cox, O. Lowry, M. Mcrorie, J.C. Martin, L. Devillers, S. Abrilian, A. Batliner, et al. The humaine database: addressing the collection and annotation of naturalistic and induced emotional data. *Affective computing and intelligent interaction*, pages 488–500, 2007.

[DD63] J.A. Deutsch and D. Deutsch. Attention: Some theoretical considerations. *Psychological review*, 70(1):80, 1963.

[DFSHB04] M.J. Dyck, C. Farrugia, I.M. Shochet, and M. Holmes-Brown. Emotion recognition/understanding ability in hearing or vision-impaired children: do sounds, sights, or

words make the difference? *Journal of Child Psychology and Psychiatry*, 45(4):789–800, 2004.

[DM80] S. Davis and P. Mermelstein. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 28(4):357–366, 1980.

[DPH00] J.R. Deller, J.G. Proakis, and J.H.L. Hansen. Discrete-time processing of speech signals. Institute of Electrical and Electronics Engineers, 2000.

[DS10] G. Dede and M.H. Sazli. Speech recognition with artificial neural networks. *Digital Signal Processing*, 20(3):763–768, 2010.

[DTU10] DTU. Java LSA Software with HAWIK Corpus Matrices, 2010. `www.imm.dtu.dk/pubdb/p.php?6320`.

[EAKK11] M. El Ayadi, M.S. Kamel, and F. Karray. Survey on speech emotion recognition: Features, classification schemes, and databases. *Pattern Recognition*, 44(3):572–587, 2011.

[Edw07] B. Edwards. The future of hearing aid technology. *Trends in Amplification*, 11(1):31–46, 2007.

[EF71] P. Ekman and W.V. Friesen. Constants across cultures in the face and emotion. *Journal of personality and social psychology*, 17(2):124, 1971.

[EH96] I. Engberg and A. Hansen. Documentation of the danish emotional speech database, 1996. `http://cpk.auc.dk/tb/speech/Emotions/`.

[EK04] J. Eggert and E. Korner. Sparse coding and nmf. In *Proceedings of IEEE International Joint Conference on Neural Networks*, volume 4, pages 2529–2533. Ieee, 2004.

[Ekm92] P. Ekman. An argument for basic emotions. *Cognition & Emotion*, 6(3-4):169–200, 1992.

[EKY+11] M. Ebisawa, M. Kogure, S. Yano, S. Matsuzaki, and Y. Wada. Estimation of direction of attention using eeg and out-of-head sound localization. In *Engineering in Medicine and Biology Society, EMBC, 2011 Annual International Conference of the IEEE*, pages 7417–7420. IEEE, 2011.

[EWG+10] F. Eyben, M. Wöllmer, A. Graves, B. Schuller, E. Douglas-Cowie, and R. Cowie. On-line emotion recognition in a 3-d activation-valence-time continuum using acoustic and linguistic cues. *Journal on Multimodal User Interfaces*, 3(1):7–19, 2010.

[EWS09] F. Eyben, M. Wollmer, and B. Schuller. Openear introducing the munich open-source emotion and affect recognition toolkit. In *3rd International Conference on Affective Computing and Intelligent Interaction and Workshops*, pages 1–6. IEEE, 2009.

[FA10] A. Frank and A. Asuncion. Uci machine learning repository, 2010.

[Fer04] R. Fernandez. *A computational model for the automatic recognition of affect in speech*. PhD thesis, Massachusetts Institute of Technology, 2004.

[Fri06] S. Frintrop. *VOCUS: A visual attention system for object detection and goal-directed search*, volume 3899. Springer-Verlag New York Inc, 2006.

[FT05] N. Fragopanagos and JG Taylor. Emotion recognition in human-computer interaction. *Neural Networks*, 18(4):389–405, 2005.

[FWG02] X. Fang, G. Wilson, and B. Giles. Subband acoustic feedback cancellation in hearing aids, 2002.

[GJ94] Z. Ghahramani and M.I. Jordan. Supervised learning from incomplete data via an em approach. In *Advances in Neural Information Processing Systems 6*. Citeseer, 1994.

[GKN08] M. Grimm, K. Kroschel, and S. Narayanan. The vera am mittag german audio-visual emotional speech database. In *Multimedia and Expo, 2008 IEEE International Conference on*, pages 865–868. IEEE, 2008.

[GP06] B. Gajic and K.K. Paliwal. Robust speech recognition in noisy environments based on subband spectral centroid histograms. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(2):600–608, 2006.

[GP10] H. Gunes and M. Pantic. Automatic, dimensional and continuous emotion recognition. *International Journal of Synthetic Emotions*, 1(1):68–99, 2010.

[GY08] M. Gales and S. Young. The application of hidden markov models in speech recognition. *Foundations and Trends in Signal Processing*, 1(3):195–304, 2008.

[HAH+92] X. Huang, F. Alleva, H.W. Hon, M.Y. Hwang, K.F. Lee, and R. Rosenfeld. *The SPHINX-II speech recognition system: an overview*. Citeseer, 1992.

[HCE+05] V. Hamacher, J. Chalupper, J. Eggers, E. Fischer, U. Kornagel, H. Puder, and U. Rass. Signal processing in high-end hearing aids: state of the art, challenges, and future trends. *EURASIP Journal on Applied Signal Processing*, 2005:2915–2929, 2005.

[HGSW04] A. Haag, S. Goronzy, P. Schaich, and J. Williams. Emotion recognition using bio-sensors: First steps towards an automatic system. *Affective Dialogue Systems*, pages 36–48, 2004.

[HHDH09] B. Han, S. Ho, R.B. Dannenberg, and E. Hwang. Smers: Music emotion recognition using support vector regression. In *Proceedings of the 10th International Conference on Music Information Retrieval (ISMIR 2009)*, 2009.

[HKM+02] V. Hozjan, Z. Kacic, A. Moreno, A. Bonafonte, and A. Nogueiras. Interface databases: Design and collection

of a multilingual emotional speech database. In *Proceedings of the Third International Conference on Language Resources and Evaluation*, pages 2019–2023, 2002.

[HL93] X. Huang and K.F. Lee. On speaker-independent, speaker-dependent, and speaker-adaptive speech recognition. *IEEE Transactions on Speech and Audio Processing*, 1(2):150–157, 1993.

[HN96] P. Horton and K. Nakai. A probabilistic classification system for predicting the cellular localization sites of proteins. In *Proceedings of the Fourth International Conference on Intelligent Systems for Molecular Biology*, pages 109–115. AAAI Press, 1996.

[Hoy02] P.O. Hoyer. Non-negative sparse coding. In *Proceedings of the 12th IEEE Workshop on Neural Networks for Signal Processing*, pages 557–565. Ieee, 2002.

[HSK+00] L.K. Hansen, S. Sigurdsson, T. Kolenda, F.A. Nielsen, U. Kjems, and J. Larsen. Modeling text with generalizable Gaussian mixtures. In *ICASSP'00. Proceedings*, volume 6, pages 3494–3497. IEEE, 2000.

[IK01] L. Itti and C. Koch. Computational modeling of visual attention. *Nature reviews neuroscience*, 2(3):194–203, 2001.

[IKN98] L. Itti, C. Koch, and E. Niebur. A model of saliency-based visual attention for rapid scene analysis. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 20(11):1254–1259, 1998.

[IMM06] IMM Technical University of Denmark. Nmf:dtu toolbox, 2006. http://cogsys.imm.dtu.dk/toolbox/nmf/index.html.

[IRT05] L. Itti, G. Rees, and J.K. Tsotsos. *Neurobiology of attention*. Academic Press, 2005.

[Itt04] L. Itti. Automatic foveation for video compression using a neurobiological model of visual attention. *Image Processing, IEEE Transactions on*, 13(10):1304–1318, 2004.

[Jac07] R. Jackendoff. Linguistics in cognitive science: The state of the art. *Linguistic review*, 24(4):347, 2007.

[Kar12] Karadogan, S. G. Emotional speech database design details and the java applet code written for the design, 2012. `www.imm.dtu.dk/pubdb/p.php?6231`.

[KBP+09a] U. Kjems, J.B. Boldt, M.S. Pedersen, T. Lunner, and D. Wang. Role of mask pattern in intelligibility of ideal binary-masked noisy speech. *The Journal of the Acoustical Society of America*, pages 1415–1426, 2009.

[KBP+09b] U. Kjems, J.B. Boldt, M.S. Pedersen, T. Lunner, and D.L. Wang. Role of mask pattern in intelligibility of ideal binary-masked noisy speech. *The Journal of the Acoustical Society of America*, 126:1415, 2009.

[KCK+10] F. Kuk, B. Crose, P. Korhonen, T. Kyhn, M. Mø rkebjerg, ML. Rank, P. Kidmose, MH. Jensen, SM. Larsen, and M. Ungstrup. Digital wireless hearing aids, part 1: A primer. *Hearing Review*, 17(3):54–67, 2010.

[KCK+11] F. Kuk, B. Crose, T. Kyhn, M. Mø rkebjerg, ML. Rank, M. Nø rgaard, and H. Pontoppidan. Digital wireless hearing aids, part 3: Audiological benefits. *Hearing Review*, 18(8):48–56, 2011.

[KHKK09] E.H. Kim, K.H. Hyun, S.H. Kim, and Y.K. Kwak. Improved emotion recognition with a novel speaker-independent feature. *IEEE/ASME Transactions on Mechatronics*, 14(3):317–325, 2009.

[KJKZ94] I. Katsavounidis, C.C. Jay Kuo, and Z. Zhang. A new initialization technique for generalized lloyd iteration. *Signal Processing Letters*, 1(10):144–146, 1994.

[KLPB10a] S.G. Karadogan, J. Larsen, M.S. Pedersen, and J.B. Boldt. Robust isolated speech recognition using binary masks. In *Proc. EUSIPCO*, pages 1988–1992, 2010.

[KLPB10b] S.G. Karadogan, J. Larsen, M.S. Pedersen, and J.B. Boldt. Robust isolated speech recognition using binary

masks. In *European Signal Processing Conference, EUSIPCO*, 2010.

[KMHL11] S.G. Karadogan, L. Marchegiani, L.K. Hansen, and J. Larsen. How efficient is estimation with missing data? In *International Conference on Acoustics, Speech and Signal Processing*. IEEE Press, 2011.

[KN07] O. Kalinli and S.S. Narayanan. A saliency-based auditory attention model with applications to unsupervised prominent syllable detection in speech. In *Eighth Annual Conference of the International Speech Communication Association*, 2007.

[KNvM98] H.J. Kappen, M.J. Nijman, and T. van Moorsel. Learning active vision. *Industrial applications of neural networks*, page 193, 1998.

[Koc10] S. Kochkin. Marketrak viii: Consumer satisfaction with hearing aids is slowly increasing. *The Hearing Journal*, 63(1):19, 2010.

[KPLL05a] C. Kayser, C.I. Petkov, M. Lippert, and N.K. Logothetis. Mechanisms for allocating auditory attention: An auditory saliency map. *Current Biology*, 15(21):1943–1947, 2005.

[KPLL05b] C. Kayser, C.I. Petkov, M. Lippert, and N.K. Logothetis. Mechanisms for allocating auditory attention: An auditory saliency map. *Current Biology*, 15(21):1943–1947, 2005.

[KRD+06] G. Keidser, K. Rohrseitz, H. Dillon, V. Hamacher, L. Carter, U. Rass, and E. Convery. The effect of multichannel wide dynamic range compression, noise reduction, and the directional microphone on horizontal localization performance in hearing aid wearers. *International Journal of Audiology*, 45(10):563–579, 2006.

[KU85] C. Koch and S. Ullman. Shifts in selective visual attention: towards the underlying neural circuitry. *Hum Neurobiol*, 4(4):219–27, 1985.

[LBC+97] P.J. Lang, M.M. Bradley, B.N. Cuthbert, et al. Motivated attention: Affect, activation, and action. *Attention and orienting: Sensory and motivational processes*, pages 97–135, 1997.

[LCCS07] E. Leon, G. Clarke, V. Callaghan, and F. Sepulveda. A user-independent real-time emotion recognition system for software agents in domestic environments. *Engineering applications of artificial intelligence*, 20(3):337–345, 2007.

[LCRD07] PA Lewis, HD Critchley, P. Rotshtein, and RJ Dolan. Neural correlates of processing valence and arousal in affective words. *Cerebral Cortex*, 17(3):742, 2007.

[LDI09] E.A.R. Losin, M. Dapretto, and M. Iacoboni. Culture in the mind's mirror: How anthropology and neuroscience can inform a model of the neural substrate for cultural imitative learning. *Progress in brain research*, 178:175–190, 2009.

[Lee90] K.F. Lee. Context-dependent phonetic hidden markov models for speaker-independent continuous speech recognition. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 38(4):599–609, 1990.

[Leo84] R. Leonard. A database for speaker-independent digit recognition. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 9, pages 328–331. IEEE, 1984.

[Lin56] D.V. Lindley. On a measure of the information provided by an experiment. *The Annals of Mathematical Statistics*, pages 986–1005, 1956.

[LN05] C.M. Lee and S.S. Narayanan. Toward detecting emotions in spoken dialogs. *Speech and Audio Processing, IEEE Transactions on*, 13(2):293–303, 2005.

[LPX+10] D. Looney, C. Park, Y. Xia, P. Kidmose, M. Ungstrup, and D.P. Mandic. Towards estimating selective audi-

tory attention from eeg using a novel time-frequency-synchronisation framework. In *Proceedings of the 2010 International Joint Conference on Neural Networks (accepted)*, 2010.

[Lun03] T. Lunner. Cognitive function in relation to hearing aid use. *International Journal of Audiology*, 42:49–58, 2003.

[LW05] Y.L. Lin and G. Wei. Speech emotion recognition based on hmm and svm. In *Machine Learning and Cybernetics, 2005. Proceedings of 2005 International Conference on*, volume 8, pages 4898–4901. IEEE, 2005.

[Mal64] N. Malebranche. 1674. recherche de la verite, 1764.

[Mar12] L. Marchegiani. *Top-Down Attention Modelling in a Cocktail Party Scenario*. PhD thesis, Sapienza University of Rome, 2012.

[Moo01] A. Moore. K-means and hierarchical clustering. *Abrufbar im Internet unter*, 2001.

[Moo10] B.C.J. Moore. The use of a loudness model to derive initial fittings of multi-channel compression hearing aids. *The Journal of the Acoustical Society of America*, 127:1846, 2010.

[Mor59] N. Moray. Attention in dichotic listening: Affective cues and the influence of instructions. *Quarterly Journal of Experimental Psychology*, 11(1):56–60, 1959.

[MR09] I.B. Mauss and M.D. Robinson. Measures of emotion: A review. *Cognition and Emotion*, 23(2):209–237, 2009.

[MRR80] C. Myers, L. Rabiner, and A. Rosenberg. Performance tradeoffs in dynamic time warping algorithms for isolated word recognition. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, 28(6):623–635, 1980.

[Mur98] K. Murphy. Hidden markov model (hmm) toolbox for matlab, 1998. http://www.cs.ubc.ca/ murphyk/Software/HMM/hmm.html.

[MVCP10] G. McKeown, M.F. Valstar, R. Cowie, and M. Pantic. The semaine corpus of emotionally coloured character interactions. In *Multimedia and Expo (ICME), 2010 IEEE International Conference on*, pages 1079–1084. IEEE, 2010.

[Neu08] A. Neuman. From the editor: Special issue on auditory scene analysis. *Trends in Amplification*, 2008.

[NFDS03] T.L. Nwe, S.W. Foo, and L.C. De Silva. Speech emotion recognition using hidden markov models. *Speech communication*, 41(4):603–623, 2003.

[NvWC+10] K. Ngo, T. van Waterschoot, M.G. Christensen, M. Moonen, S.H. Jensen, and J. Wouters. Adaptive feedback cancellation in hearing aids using a sinusoidal near-end signal model. In *IEEE International Conference on Acoustics Speech and Signal Processing*, pages 181–184. IEEE, 2010.

[OF06] R.C. O'Reilly and M.J. Frank. Making working memory work: a computational model of learning in the prefrontal cortex and basal ganglia. *Neural Computation*, 18(2):283–328, 2006.

[O'S08] D. O'Shaughnessy. Invited paper: Automatic speech recognition: History, methods and challenges. *Pattern Recognition*, 41(10):2965–2979, 2008.

[Pas99] H.E. Pashler. *The psychology of attention*. The MIT Press, 1999.

[PH10a] M.K. Petersen and L.K. Hansen. Modeling lyrics as emotional semantics. *Proceedings of YoungCT, KAIST Korea Advanced Institute of Science and Technology*, 2010.

[PH10b] P.C. Petrantonakis and L.J. Hadjileontiadis. Emotion recognition from eeg using higher order crossings. *IEEE Transactions on Information Technology in Biomedicine*, 14(2):186–197, 2010.

[PHA92] R.D. Patterson, J. Holdsworth, and M. Allerhand. Auditory models as preprocessors for speech recognition. *The*

Auditory Processing of Speech: from Auditory Periphery to Words, pages 67–89, 1992.

[Phi06] D. Phillips. *Preparation Course for the TOEFL: Next Generation IBT*. Longman, 2006.

[PM10] D.L. Poole and A.K. Mackworth. *Artificial Intelligence: foundations of computational agents*. Cambridge Univ Pr, 2010.

[PMKT11] H. Peng, D. Majoe, and T. Kaegi-Trachsel. Design and application of a novel wearable eeg system for e-healthcare. In *Proceedings of 2011 international workshop on Ubiquitous affective awareness and intelligent interaction*, pages 1–8. ACM, 2011.

[Pos80] M.I. Posner. Orienting of attention. *Quarterly journal of experimental psychology*, 32(1):3–25, 1980.

[Pos11] M.I. Posner. *Cognitive neuroscience of attention*. The Guilford Press, 2011.

[PSB+10] A. Prasad, Y. Srinivas, P. Brahmaiah, et al. Gender based emotion recognition system for telugu rural dialects using hidden markov models. *Arxiv preprint arXiv:1006.4548*, 2010.

[PY03] O. Pierre-Yves. The production and recognition of emotions in speech: features and algorithms. *International Journal of Human-Computer Studies*, 59(1):157–183, 2003.

[Rab89] L.R. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.

[Rad] The Danish Radio. The drcd sound effects library.

[RFPC07] J. Redondo, I. Fraga, I. Padrón, and M. Comesaña. The spanish adaptation of anew (affective norms for english words). *Behavior research methods*, 39(3):600–605, 2007.

[RFST+08] M. Rudner, C. Foo, E. Sundewall-Thoren, T. Lunner, and J. Ronnberg. Phonological mismatch and explicit cognitive processing in a sample of 102 hearing-aid users. *International Journal of Audiology*, 47(Supplement 2):91–98, 2008.

[RGH+09] D. Rybach, C. Gollan, G. Heigold, B. Hoffmeister, J. Lööf, R. Schlüter, and H. Ney. The rwth aachen university open source speech recognition system. In *Tenth Annual Conference of the International Speech Communication Association*, 2009.

[Ric05] T.A. Ricketts. Directional hearing aids: Then and now. *Journal of Rehabilitation Research and Development*, 42(4):133, 2005.

[Rob01] David Robinson. Replay gain, 2001. `http://www.replaygain.org/`.

[Ros03] S.M. Ross. Peirce's criterion for the elimination of suspect experimental data. *Journal of Engineering Technology*, 20(2):38–41, 2003.

[Rub76] D.B. Rubin. Inference and missing data. *Biometrika*, 63(3):581, 1976.

[SC11] G. Saon and J.T. Chien. Bayesian sensing hidden markov models for speech recognition. In *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*, pages 5056–5059. IEEE, 2011.

[SCG+10] S. Scanzio, S. Cumani, R. Gemello, F. Mana, and P. Laface. Parallel implementation of artificial neural network training for speech recognition. *Pattern Recognition Letters*, 31(11):1302–1309, 2010.

[Sch01] T. Schneider. Analysis of incomplete climate data: Estimation of mean values and covariance matrices and imputation of missing values. *Journal of Climate*, 14(5):853–871, 2001.

[Sch07] O. Scharenborg. Reaching over the gap: A review of efforts to link human and automatic speech recognition research. *Speech Communication*, 49(5):336–347, 2007.

[Sch08] A. Schaub. *Digital hearing aids*. Thieme Medical Pub, 2008.

[Sch11] B. Schuller. Recognizing affect from linguistic information in 3d continuous space. *IEEE Transactions on Affective Computing*, (99):1–1, 2011.

[SD07] T. Su and J.G. Dy. In search of deterministic methods for initializing K-means and Gaussian mixture clustering. *Intelligent Data Analysis*, 11(4):319–338, 2007.

[SG02] J.L. Schafer and J.W. Graham. Missing data: our view of the state of the art. *Psychological methods*, 7(2):147, 2002.

[SGB01] M. Sarter, B. Givens, and J.P. Bruno. The cognitive neuroscience of sustained attention: where top-down meets bottom-up. *Brain Research Reviews*, 35(2):146–160, 2001.

[SJ01] J.A. Sloboda and P.N. Juslin. *Psychological perspectives on music and emotion*, pages 71–104. Oxford University Press, 2001.

[SLH07] M.N. Schmidt, J. Larsen, and F.T. Hsiao. Wind noise reduction using non-negative sparse coding. In *IEEE Workshop on Machine Learning for Signal Processing*, pages 431–436. IEEE, 2007.

[SMJ07] R.A. Stevenson, J.A. Mikels, and T.W. James. Characterization of the affective norms for english words by discrete emotional categories. *Behavior research methods*, 39(4):1020–1024, 2007.

[SRL04] B. Schuller, G. Rigoll, and M. Lang. Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine-belief network architecture. In *Acoustics, Speech, and Signal Pro-*

cessing, 2004. Proceedings.(ICASSP'04). IEEE International Conference on, volume 1, pages I–577. IEEE, 2004.

[SRM+05] B. Schuller, S. Reiter, R. Muller, M. Al-Hames, M. Lang, and G. Rigoll. Speaker independent speech emotion recognition by ensemble classification. In *IEEE International Conference on Multimedia and Expo*, pages 864–867. IEEE, 2005.

[SS04] A.J. Smola and B. Schölkopf. A tutorial on support vector regression. *Statistics and computing*, 14(3):199–222, 2004.

[SUMIGA+07] R. Solera-Urena, D. Martín-Iglesias, A. Gallardo-Antolín, C. Peláez-Moreno, and F. Díaz-de María. Robust asr using support vector machines. *Speech Communication*, 49(4):253–267, 2007.

[TE09] M. Tinston and Y. Ephraim. Speech enhancement using the multistage wiener filter. In *43rd Annual Conference on Information Sciences and Systems, 2009. CISS 2009*, pages 55–60. IEEE, 2009.

[TG80] A.M. Treisman and G. Gelade. A feature-integration theory of attention. *Cognitive psychology*, 12(1):97–136, 1980.

[Tre60] A. M. Treisman. Contextual cues in selective listening. *Quarterly Journal of Experimental Psychology*, 12:242–248, 1960.

[Ung00] S.K.L.G. Ungerleider. Mechanisms of visual attention in the human cortex. *Annual review of neuroscience*, 23(1):315–341, 2000.

[Uni02] University of Pennsylvania Linguistic Data Consortium. Emotional prosody speech and transcripts, 2002. `http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2002S28S`.

[VA05] T. Vogt and E. André. Comparing feature sets for acted and spontaneous speech in view of automatic emotion

recognition. In *IEEE International Conference on Multimedia and Expo*, pages 474–477. IEEE, 2005.

[Ver06] Ververidis, D. and Kotropoulos, C. Emotional speech recognition, resources, features, and methods. *Speech communication*, 48(9):1162–1181, 2006.

[VK05] D. Ververidis and C. Kotropoulos. Emotional speech classification using gaussian mixture models and the sequential floating forward selection algorithm. In *IEEE International Conference on Multimedia and Expo*, pages 1500–1503. IEEE, 2005.

[VTR92] F.J. Varela, E. Thompson, and E. Rosch. *The embodied mind: Cognitive science and human experience*. The MIT Press, 1992.

[Wan05a] D. Wang. On ideal binary mask as the computational goal of auditory scene analysis. *Speech separation by humans and machines*, pages 181–197, 2005.

[Wan05b] D.L. Wang. On ideal binary mask as the computational goal of auditory scene analysis. *Speech separation by humans and machines*, pages 181–197, 2005.

[WB06] D.L. Wang and G.J. Brown. Computational auditory scene analysis: Principles, algorithms, and applications. *Recherche*, 67:02, 2006.

[WC95] N. Wood and N. Cowan. The cocktail party phenomenon revisited: how frequent are attention shifts to one's name in an irrelevant auditory channel? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 21(1):255, 1995.

[WKP+08] D.L. Wang, U. Kjems, M.S. Pedersen, J.B. Boldt, and T. Lunner. Speech perception of noise with binary gains. *The Journal of the Acoustical Society of America*, 124:2303, 2008.

[WKP+09] D.L. Wang, U. Kjems, M.S. Pedersen, J.B. Boldt, and T. Lunner. Speech intelligibility in background noise with

ideal binary time-frequency masking. *The Journal of the Acoustical Society of America*, 125:2336, 2009.

[YNP11] S. Yildirim, S. Narayanan, and A. Potamianos. Detecting emotional state of a child in a conversational computer game. *Computer Speech & Language*, 25(1):29–44, 2011.

[You93] S.J. Young. The htk hidden markov model toolkit: Design and philosophy. *Department of Engineering, Cambridge University, UK, Tech. Rep. TR*, 153, 1993.

[YSL07] C. Yang, F.K. Soong, and T. Lee. Static and dynamic spectral features: Their noise robustness and optimal weights for asr. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(3):1087–1097, 2007.

[YTCZ11] C. Yu, Q. Tian, F. Cheng, and S. Zhang. Speech emotion recognition using support vector machines. *Advanced Research on Computer Science and Information Engineering*, pages 215–220, 2011.

[ZKK+09] A.A. Zekveld, S.E. Kramer, J.M. Kessens, M.S.M.G. Vlaming, and T. Houtgast. The influence of age, hearing, and working memory on the speech comprehension benefit derived from an automatic speech recognition system. *Ear and hearing*, 30(2):262, 2009.

[ZZL11] S. Zhang, X. Zhao, and B. Lei. Spoken emotion recognition using radial basis function neural network. *Advances in Computer Science, Environment, Ecoinformatics, and Education*, pages 437–442, 2011.