M 2014

U.PORTO

FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

FEATURE SETS FOR STRESSED SPEECH DISCRIMINATION

MARIANA DIMAS JULIÃO

DISSERTAÇÃO DE MESTRADO APRESENTADA
À FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO EM
ENGENHARIA DA INFORMAÇÃO

FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO

Feature Sets for Stressed Speech Discrimination

Mariana Dimas Julião

U.PORTO

FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

Mestrado em Engenharia da Informação

Supervisor: Ana Aguiar

Second Supervisor: Aníbal Ferreira

December 18, 2014

Feature Sets for Stressed Speech Discrimination

Mariana Dimas Julião

Mestrado em Engenharia da Informação

December 18, 2014

Resumo

O *stress* emotivo é frequentemente sentido durante o discurso em público, e é uma das fobias mais frequentes em adultos. Este factor manifesta-se normalmente através formas detectáveis por humanos – o que acontece tão mais facilmente quanto mais conhecido do ouvinte for o orador –, o que pode querer dizer que o sinal de voz é afectado pelo stress de formas mais ou menos previsíveis. Por esta razão, é legítimo supor que que seja possível treinar um computador para que aprenda a fazer esta distinção.

Muitos estudos foram já feitos sobre reconhecimento automático de emoções. No entanto, devido à vastidão do seu âmbito relativamente às emoções a reconhecer, ao número de oradores, e aos classificadores, não é claro quais são as características quep pdem ser as mais eficientes na tarefa de distinção de *stress*. Em particular, não é claro quais são as características que podem ser mais eficientes para os nossos dadas. Aqui, remetemo-nos à distinção de *stress*, para 15 oradores. Tentamos perceber quais dos subconjuntos propostos do conjunto inicial de 6365 características, incluindo funcionais de características e características baseadas no Teager Energy Operator, melhor contibuem para esta classificação.

Os dados são recolhidos através duma abordagem ecológica, não qual o *stress* não é induzido, mas está naturalmente presente no sistema. Dados fisiológicos são gravados, a fim de permitirem a futura anotação dos segmentos de discurso. Esta anotação é feita binariamente, na medida em que os segmentos de discurso são apenas classificados como “com *stress*” ou “neutro’.

Abstract

Emotional stress is often experienced during public speaking, and is one of the most common phobias in adults. This can usually be detected by the human listener – and this is as easy as the speaker is known by the listener –, which can mean that speech signal is affected in predictable ways by stress. For this reason, it must be possible to teach a machine how to discriminate it.

Many studies have already been done in automatic emotion recognition. However, due to their broadness of scope, concerning number of emotions to recognize, number of speakers, and types of features, it is not clear which features can be the most efficient for the task of stress discrimination. In particular, it is not clear which features can be the most efficient for our data set. Here we address the problem of stress discrimination, for 15 speakers. We try to understand which of the proposed subsets of the inicial set of 6365 features, including functional features and Teager Energy Operator -based features, enable a better discrimination.

Data is collected using an environmental approach, where stress is not induced but already present in the scenario. Physiologic data is recorded, in order to provide for further annotation of speech utterances. Stress annotation is only taken binarily, meaning that the speech utterance is either annotated as “stressed” or as “neutral”.

Agradecimentos

Tenho a agradecer, antes de mais, à minha orientadora, Prof. Ana Aguiar, por tudo o que pude aprender durante o projecto, bem como ao meu co-orientador, Prof. Aníbal Ferreira, pela sua disponibilidade e pelo seu interesse pelo meu trabalho e pelo meu progresso.

Todas a equipa do VOCE merece um grande agradecimento meu, principalmente: Prof. Ana Aguiar, Traian Abrudan, Jorge Silva, Jaime Ferreira, Hugo Meinedo e, mais recentemente, Helena Moniz e Fernando Batista. Agradeço muito à equipa no INESC-ID pela sua colaboração neste trabalho. Gostava também de agradecer ao Ricardo Sousa, pelas suas sugestões para o Informed Choice Feature Set, bem como ao Pedro Santos e ao João Rodrigues, do IT, pelo acolhimento e pelo apoio.

Do DEEC tenho a agradecer a preciosa colaboração de José António Nogueira e de Rosário Macedo, nas partes menos académicas do processo.

Quero muito agradecer a cada um dos professores que encontrei no MeInf – pela disponibilidade que sempre tiveram para me ajudar a aprender mais. E aos meus colegas, claro.

Dou graças à (e pela) minha família: Mãe, Pai, Pedro, e todos – por me deixarem falhar, e falhar, todas as vezes. E por continuarem a esperar que – um dia – eu falhe melhor.

De entre os maravilhosos grupos de pessoas que tive a fortuna de encontrar no Porto, há um que me merece especial atenção neste local. Agradeço aos meus amigos que são a Sociedade de Debates da Universidade do Porto, por aproximarem a Universidade do que sempre deveria ser. E nomeio, em especial, os seus sócios-fundadores, que tenho o enorme privilégio de ter por amigos: Joana, Luísa, Ary – que bom é ter-vos tão perto!

Por fim, João. Ao João agradeço por todas as maravilhas que me trouxe. Além daquelas para cujo agradecimento não é este o local, tenho a agradecer os pertinentes lembretes relativos ao quanto gosto do que faço,

e o ter-me transmitido que mesmo tudo o que se possa aprender na faculdade nunca há-de chegar para as exigências do mundo lá fora.

Mariana

Contents

List of Figures

List of Tables

Abreviaturas e Símbolos

CS	Complete Set
IC	Informed Choice
FS	Feature Set
LLD	Low Level Descriptor
RR	Recognition Rate
SER	Speech Emotion Recognition
SVM	Support Vector Machine
TEO	Teager Energy Operator

Chapter 1

Introduction

1.1 Motivation

According to Scherer et al. (2002) it was in the beginning of the XX century that scientists started trying to identify the acoustic profiles of the major emotions, and it was in the 50s and 60s – when stress became a matter of curiosity – that identification of acoustic correlates began. This subject then began to be looked at with more regularity from the 80s on. The reasons to develop automated ways of speech classification concerning stress are several-fold, since it is a nonintrusive way to detect stress: ranking emergency calls (Demenko (2008)), improve speech recognition systems for it is know that environmentally induced stress leads to fails on speech recognition systems – Hansen et al. (1998), and in our case providing the speaker with tips on how to improve their public speaking, as happens with VOCE (Aguiar et al. (2014)).

Stress can have several manifestations that allow people to perceive stress in others. These manifestations can happen v.g. through facial expressions, body posture, or voice. Concerning the relation between the size and quality of the training set of a classifier and the performance of the classification, classification in humans is not specially different from the one in artificial media. Consequently, the greater the number of times one person has been with another, the greater the accuracy in the classification of their emotional state. Although humans automatically have input from different sources, as vision and audition of both words and intonation, often emotion discrimination can be done using only one of those.

Emotion recognition systems try to model the processes in human brain that enable emotion recognition. VOCE belongs to this group. Here, we assume that stress recognition is possible from speech alone. And then, we try to understand which speech properties are most relevant for that guess. To be more precise, we try to understand which combinations of speech properties are most relevant for that guess.

1.2 Context: VOCE

This thesis falls within the scope of the VOCE project, which aims at developing algorithms to identify stress from live speech. To do so, speech is recorded in an ecological environment, with physiological validation. Having an ecological approach means that stress is not induced on the speakers, it is rather an expected consequence of the environment the speaker is in. This environment corresponds to a public speech event that occurs within academic context, as presentations of coursework or research seminars. The process of data collecting is detailed in Aguiar et al. (2013) and VOCE corpus is presented in Aguiar et al. (2014).

The speech is segmented into utterance-like units, which are annotated using physiological signals. These sentence-like units are henceforth named utterances. The annotation is based on the mean heart rate. For each speaker, the utterances annotated as stressed are the ones with a mean heart rate value belonging to the higher quartile. The remaining ones are labelled as neutral.

The ultimate goal of VOCE is to develop an application than can detect stress in live speech, while giving the user feedback on the quality of their speech in real time. For this reason, these algorithms must be fast. Stress detection requires feature extraction and classification. By reducing the feature set, there is a reduction on the amount of features that need to be extracted and, consequently, also the complexity of the classifier decreases. Eventually, feature selection increases classifier accuracy. This is due to the trade off there is between complexity and precision, expressed as delay and resource consumption in a potentially limited device. These small feature sets are then chosen based on the ability for stress discrimination of their features together.

1.3 Overview

As briefly introduced in Section 1.1, the objective of this thesis is to understand which of the available feature sets provides the best performance for stressed/neutral discrimination. This process is done by comparing the results of the performance of a classifier when given the different feature sets. By performance we mean the rate of correctly classified utterances.

Two feature sets are extracted initially from the audios belonging to this corpus. These sets are Feature Set I: the group of functional features, and Feature Set II: the group of Teager Energy Operator-Based Features, to be detailed on Chapter 4. The union of these two sets corresponds to all the features we are considering in this work. To obtain subsets of these sets, we firstly considered a subset of Feature Set I: Informed Choice. This subset is based on the answers we got from experts in speech processing when asked about the feature sets they would guess to have the best performance in this task.

Starting from these feature sets, feature selection algorithms ran on each set and on combinations of feature sets to reduce the dimensionality of the classification problem. With this process we obtained feature subsets. The results of classification of these subsets are compared to the set from which they were part, as well as to other resulting feature sets.

1.4 Structure

This document is structured as follows:

Chapter 2, describes the work that has already been done in this field.

Then, Chapter 3, describes what this thesis is about: the problem it addresses, the methodology, what its contribution is.

On Chapter 4, the features that are considered in this work are presented. The way features are computed, as well as their main known properties are also explained.

Then there is a chapter for the results and discussion, Chapter 5. Finally, Conclusions and Future Work, correspond to Chapter 6, where the perspective of the authors on the goal achievement is stated, and where it is described how this work shall be continued on the future.

Chapter 2

Related Work

By exposing the related work, this section presents and clarifies the main concepts in this thesis. It starts by introducing the concept of stress, the ways it can arise, and ways of recording it. Since this work shares insight with other fields of artificial intelligence that are also related to speech, some differences and similarities between them are exposed. This section also shows datasets and their properties, criteria concerning speech unit choices, and means for stress annotation. The most frequent types of features are announced. Features based on Teager Energy Operator are given special attention for their extraction is part of this work. Finally, this section states the most relevant methods for feature selection and classification.

2.1 Stress

2.1.1 Definition

Stress is very hard to define, and so has been speech under stress. Hansen (1996) states that "Stress is a psychological state that is response to a perceived threat or task demand and is normally accompanied by specific emotions.". Lu et al. (2012) say that “in general terms, stress is the reaction of an organism to a change in its equilibrium”. According to the same authors, one of the issues concerning stress classification is the fact that it is very hard to isolate emotions, and stress usually comes in combination with other emotions, like anger or sadness. These authors even state that because of that “Stress detection in speech is therefore often wrapped up in speech emotion detection.” In this work it is also stated that the main difference between stress detection and speech emotion recognition is that it is easier to link a stressor to the stress it causes than the cause of an emotion to its result. Hansen, one of the most prolific authors in the subject, goes to the point of saying in Hansen and Patil (2007) that “any deviation in speech with respect to neutral style, whether it is speaking style, word selection, word usage, sentence duration is termed as speech under stress” and that, therefore, "speech under stressful conditions refers to speech spoken under some environmental factor or emotional state which perturbs speech production from a natural, conversational framework".

2.1.2 Speech Recognition, Speech Synthesis, Speech Discrimination

There are several fields of study that are very close to our own, even due to the use of the same tools. These are speech recognition and speech synthesis. Much of the information of the features and of the processes used here can be found in the literature corresponding to those topics.

Speech synthesis is the process of artificially producing speech close to human speech. Adding an emotional component to speech, as certain intonations, is helpful to make it sound natural (Bou-Ghazale and Hansen (1996)). Also, Bou-Ghazale and Hansen (1998), propose a method that consists of modeling the variation "in pitch contour, voiced speech duration, and average spectral structure". Ruzanski et al. (2005) even establish a connection between stress and vowel duration, which can also be taken as a parameter to add to speech synthesis. These features, however, shall be described in Section 2.5.

Stress has a great impact in speech recognition, in the sense that the performance of recognition algorithms severely degrades in the presence of stress (Bou-Ghazale and Hansen (2000), Hansen (1996)). According to Bou-Ghazale and Hansen (2000), there are three main ways of dealing with this problem: through robust features, stress equalization methods, and model adjustment or training methods. The referred study makes an evaluation of the effectiveness of previously proposed features and new noise robust features are proposed. Jabloun et al. (1999) also suggest features for speech recognition in car noise – which have also been used for stressed speech discrimination.

Speech discrimination is the task of discriminating whether the speaker was experiencing stress while speech was made. In terms of process, this is very close to Speech Emotion Recognition, if not a part of Speech Emotion Recognition. Ruiz et al. (1996) refer that there are significant differences concerning what is said to be stress, because either "sudden inhalation of ammonia" or "the evidence of imminent death in an aircraft crash" are taken as stressful situations, meaning that care must be taken when comparing studies. Moreover, most of articles focus on one kind of stressful situation only. This is also the approach in this work, and VOCE.

2.1.3 Different "Types" of Stress

Works on speech under stress address different kinds of stress, which can be categorised into physical, emotional, and cognitive. These kinds of stress may be deeply related to the way they are experimentally induced, but they can exist naturally.

Physical stress, according to Lu et al. (2012), is the consequence of a threat to an individual's physical equilibrium. Several studies address the problem of the loss of quality induced in the speech signal, which tend to be close to the effects of emotional or cognitive stress. Experienced G-force in aircraft cockpits is an example of this, widely used in Hansen et al. (1998).

Emotional stress happens when the speaker is in a stressful situation where they feel some threat or insecurity. Lu et al. (2012) say that it comes from a threat to a person's self-esteem. This can range from aircraft communications to emergency calls and to academic presentations (Aguiar et al. (2013), Yang et al. (2012), Scherer (2003), Ververidis and Kotropoulos (2006)).

Cognitive stress is the consequence of some intellectually demanding tasks the speaker has to do. Scherer et al. (2002) and Fernandez and Picard (2003) show examples of this.

2.1.4 Types of Recordings

Studies on emotion or stress recognition done to date have relied on different approaches to obtain speech in stressful situations. Ruiz et al. (1996) considers two different types: real events and artificial laboratory situations. However, artificial laboratory situations can be divided in acted and induced. Therefore, we consider here three main approaches: environmental, acted, and induced.

In the beginning of speech emotion recognition (SER), data came from actors that acted certain emotions (Williams and Kenneth (1972), Vogt et al. (2008)). This may have some advantages, like clearness or emotion separability, but it is not a realistic approach. Actors are trained to represent in an identifiable manner: a way that exaggerates what happens in real life, in order to be properly understood by people standing a few meters away.

Induced stress exists when researchers make the speakers experience stress in someway. According to Ruiz et al. (1996), one of the first ways of stress induction may have been audio-feedback perturbation, where speakers may have had to strain their voices due to the noise in the environment. Then, there are cognitive stressors: some cognitive tasks that induce stress on the speaker, as doing several cognitive demanding tasks at the same time, or doing difficult computations. It happens in Fernandez and Picard (2003), for instance, where drivers have a cognitive load by having to answer mental challenges while driving. Another example is in Scherer et al. (2002), where speakers had to perform logical deductions and, given two phone rings, respond to only one, while keeping the logical deduction.

Environmental stress happens when stressors are in the environment where the recording is carried out, without being artificially produced by the researchers. It is what happens in Lu et al. (2012), where physiological data from speakers is recorded along with the audio, in a certain period of time. It is also what we can see in Demenko (2008), in which emergency calls are recorded. Most of the work done inside this category is related to aeronautics, due to stress induced by dangers. SUSAS – Speech Under Simulated and Actual Stress Database – Hansen et al. (1998) is an example of this, for the great majority of its recordings are not Simulated Stress (some of the recordings are acted). Schuller et al. (2011) state that the context of recording speakers unaware of the recording is not only ideal, but also frequently possible. However, it has the drawback that only rarely that data can be freely distributed, due to privacy reasons. Therefore, to set an environment as close to natural as possible shall be the intent. The VOCE corpus used in this work and described in Aguiar et al. (2014) falls in this category.

Some meta-information is commonly helpful for stressed speech discrimination, in the sense that speech signal alone is not the major provider of information for human listeners (Douglas-Cowie et al. (2003)). These authors also note the cultural nuances of stress, and the way some voice intonations are perceived differently, according to the culture in which they exist. No such information is available in the VOCE corpus.

2.2 Datasets

Most of the works referenced in this chapter are based on available speech datasets, which tend to have very different constitutions, both in number of speakers and of utterances, but also in the emotional range covered. Douglas-Cowie et al. (2003) and Ververidis and Kotropoulos (2003) provide extensive information on the available databases at the time they were written. In this section, we address databases that were used in articles that made the literature review of this work.

Of all available databases, SUSAS – Speech Under Simulated and Actual Stress – (Hansen et al. (1998)) is the database with higher visibility, according to this research. Many studies were based on its material, as Zhou et al. (2001), Hansen et al. (2011), Bou-Ghazale and Hansen (1996), Bou-Ghazale and Hansen (1998), Hansen and Patil (2007), and Sarikaya and Gowdy (1998). It has five domains: Talking Styles, Single Tracking Task, Dual Tracking Task, Actual Speech Under Stress, and Psychiatric Analysis, from a total of 32 speakers. The first four correspond to recordings of 35 different words from aircraft communication. The total number of utterances is 16.000.

The database Fernandez and Picard (2003), named in Ververidis and Kotropoulos (2003) as "Database 11", as described in the article, is made of 598 utterances, corresponding to four different speakers. The subjects were asked to answer mathematical questions while driving trucks. There were two speeds: 60 m.p.h. and 120 m.p.h.. While in the first the subject was asked an answer every 9 seconds, in the second the subject was asked to answer every 4 seconds.

Demenko (2008), and Demenko and Jastrzebska (2012) use a database of emergency calls of Poznan Police, made of "spontaneous speech recordings that consists of crime/offence notification and police intervention requests". Forty-five speakers were chosen after acoustical evaluation. Preliminary annotation was done by students trained in phonetics, and included description of type of dialogue, characteristics of the speech ("suprasegmental features"), context, background acoustics. In order to study stress, two utterances from situations of different arousals were considered for each speech.

The Danish Emotional Database is described in Engberg and Hansen (1996), and used in Lin and Wei (2005) and Ververidis and Kotropoulos (2005). Two male actors and two female actors that are familiar with radio features act five different emotional states: anger, happiness, neutral, sadness, and surprise.

The intelligibility of speech in people after advanced head and neck cancer treated with chemoradiotherapy is the subject of Clapham et al. (2012). It has a database to allow for the development of automatic evaluation of intelligibility in those patients, to avoid drawbacks of human listening perception.

2.3 Speech Units

Concerning the size of speech segments to analyse, some comparisons have already been done.

Vogt et al. (2008) says that a good unit for studying speech emotion recognition shall be: "(1) long enough to reliably calculate features by means of statistical functions and (2) short enough to guarantee stable acoustic properties with respect to emotions within the segment". Short-term and long-term are two types of feature approach, and there are no certainties yet on which is better. Short-term features are computed within short-time frames of 10 or 20 ms; long-term features are computed for the whole unit, being usually statistics on the short-time features. Some studies – namely those based on SUSAS – also use words or syllables "tokens" as units.

Fernandez and Picard (2003) states that, concerning speech recognition, the most suitable time scale is the one of the phonemes. For stress discrimination, it proposes global-utterance level features, and says that statistics of each feature along the utterance can provide a good utterance-level representation. They postulate that since they are interested in modelling supra-linguistic phenomena, a time-scale rougher that the phonemes must be appropriate.

In parallel to this, Ververidis and Kotropoulos (2005) adds his experimental results, saying that utterances from isolated words are richer in emotions than those inside complete sentences. This is directly connected to the notion of supra-linguistic feature, once people tend to say more isolated and repeated words when nervous. Furthermore, even in neutral speech, isolated words (v.g. as interjections) tend to be more emphatic than non-isolated words.

Works as Lu et al. (2012), Hansen et al. (2011), and Zhou et al. (2001) have used short-term features, while Ververidis and Kotropoulos (2005),Batliner et al. (2006), and Fernandez and Picard (2003) have used long-term features.

In this work we benefit from previous works on speech segmentation, both at the utterance level and at the phoneme level, being the utterance a sentence-like unit. These are presented at Batista et al. (2012) and Abad (2012).

2.4 Annotation

The way annotation of speech is done also varies in literature. Many studies consider speech as stressed as long as the speaker is in a stressful situation, also called block annotation. This is what happens in Fernandez and Picard (2003), Zhou et al. (2001), or in any work relying on SUSAS database Hansen et al. (1998).

Self-evaluation of the speakers, either through a simple description or through answering a questionnaire after speaking, was used in Scherer et al. (2008) and Chen and Li (2013).

Other studies use human listeners (usually experts) to do the annotation, as Minematsu et al. (2002), Banse and Scherer (1996).

Physiological data is used in Lu et al. (2012) and Aguiar et al. (2014). In Lu et al. (2012), galvanic skin resistance is used, as skin conductance tends to increase with arousal. Aguiar et al. (2014) use heart rate variability, although a questionnaire is also answered by the speaker). Sharma and Gedeon (2012) present a review of how physiological data can be used to annotate stress.

Schuller et al. (2011) stresses the importance of annotation done through combination of data from several media. Concerning this authors mention not only audiovisual information, but also input about context.

2.5 Features for Emotion Recognition

The most consensual feature for stress discrimination has been the pitch, or fundamental frequency (F0). Then, several metrics for energy and formant changes are proposed. Sometimes, information of energy and formants is represented with Mel-Frequency Cepstral Coefficients (MFCCs) (Zhou et al. (2001), Sarikaya and Gowdy (1998), Demenko (2008)). After these, came frequency and amplitude perturbations – Jitter and Shimmer –, and other measures of voice quality, like Noise to Harmonics Ratio and Subharmonics to Harmonics Ratio (Vogt et al. (2008), Sun (2000)). Teager Energy Operator-based features have also shown to perform well in speech under stress (Zhou et al. (2001)), and this work gives them special emphasis.

Furthermore, Vogt et al. (2008) state the importance of supra-segmental acoustic phenomena that can be taken as global emotion features, although not as extractable as others mentioned so far: "hyper-clear speech, pauses inside words, syllable lengthening, off-talk, respiration disfluency cues, inspiration, expiration, mouth noise, laughter, crying, unintelligible voice. Though these have been mainly annotated by hand, automatic extraction would also be possible in some cases."

Schuller et al. (2011) agrees with the previous vision, since it divides features into two major groups: acoustic and linguistic. This separation is fundamentally technology-based, since they are extracted in very different ways. It is also said that the contributions of these types of features strongly depend on the application

In other work, the same author, Schuller et al. (2008), explains functional features, saying that statistical noise due to extrema can be overcome by considering hierarchical statistical functional over small units as words or chunks: "a whole turn of speech". This approach leads to a reduction of information. However, at the same time, it allows for generalisation due to loss of information on speaker content.

2.5.1 Teager Energy Operator

After finding non-linearities in the mechanisms of speech production, Teager (Teager (1980)) realised that describing these non-linearities through only linear operators would be a rough work (Teager and Teager (1985), and Teager and Teager (1990)). The development of the improved algorithm, however, was brought by Kaiser (who named it after Teager), in Kaiser (1990a) and generalized to continuous signals in Kaiser (1990b). This algorithm has been applied to several problems in several fields, from speech to image processing (Kvedalen (2003)).

The Teager Energy Operator (TEO) is defined as

$$\Psi_c[x(t)] = \left(\frac{d}{dt}x(t)\right)^2 - x(t)\left(\frac{d^2}{dt^2}x(t)\right) \tag{2.1}$$

$$= [\dot{x}(t)]^2 - x(t)\ddot{x}(t) \tag{2.2}$$

where $\Psi[.]$ is the Teager energy operator, and $x(t)$ is a single component of the continuous speech signal.

Its form for discrete-time signals is

$$\Psi[x(n)] = x^2(n) - x(n+1)x(n-1) \quad . \tag{2.3}$$

Zhou and his collaborators came up with three features based on TEO that were claimed to allow classification of speech under stress (Zhou et al. (2001)). These are Variation of FM component, Normalized TEO autocorrelation envelope, and Critical band based TEO autocorrelation envelope. Many further studies rely on these features – namely those conducted by J.H.L.Hansen (v.g. Hansen et al. (2011), Lu et al. (2012)). Jabloun, on his turn, came up with TEO-based cepstal coefficients (Jabloun et al. (1999)), that showed to improve speech recognition in car noise. These TEO-based cepstal coefficients were afterwards interestingly used by Fernandez and Piccard (Fernandez and Picard (2003)).

In Lu et al. (2012), TEO-based features reached the best performance for stressed speech discrimination outdoor, but not indoor. They also have been used to do voiced-unvoiced classification (N. et al. (2003)). In the latter work, the advantages of TEO are enunciaded: because only three samples are needed for the energy computation at each time instant, it is nearly instantaneous. Therefore, this time resolution allows to capture energy fluctuation, and also a robust AM-FM estimation in noisy environment. Cairns et al. (1996), use Teager Energy Operator in the development of a system for hypernasal speech detection.

2.6 Feature Selection

Concerning feature selection in general, Guyon and Elisseeff (2003) provide us with a large review on the methods for feature selection. It starts by explaining the advantages feature selection may bring: "facilitating data visualisation and data understanding, reducing the measurement and storage requirements, reducing training and utilisation times, defying the curse of dimensionality to improve prediction performance.". For the present work, we are interested in the last two points. Authors clearly state that their work is more about feature set selection than feature selection, in order to build a good predictor. This is also clearly explained in their section for "Small but Revealing Examples", where the meaning of the previous sentence is shown: often happens that well-ranked individual features do not provide good results, as well as it happens that apparently useless become useful when joining others. For feature subset selection, Guyon and Elisseeff (2003) expose three main methods: wrappers, filters, and embedded methods.

Kira and Rendell (1992) proposes the Relief algorithm for feature selection, which is good at selecting relevant features, but that can easily choose redundant features too, against which Guyon and Elisseeff (2003) advised.

Wrappers for feature set selection do not aim at finding the best subset possible, but they greedily combine features until they find a local maximum of performance. These algorithms were introduced in Kohavi and John (1997), and are explained in Section 4.2.

Among the works on speech emotion recognition, Banse and Scherer (1996) and Demenko and Jastrzebska (2012) use discriminant analysis, and Schuller et al. (2008) uses wrapper.

2.7 Classifiers

According to Vogt et al. (2008), and as one can easily see, there is an intrinsic relation between the type of classifiers and features: for global statistic features, a "static" classifier – like support vector machine (SVM) – should be used; while for short-term features a dynamic classifier such as a hidden markov model (HMM) should be used. It also says that while in the first part the dynamic properties of emotions are captured by the features, in the second they are dealt with by the classifier.

Ververidis and Kotropoulos (2006) and Vogt et al. (2008) state the intrinsic relation there is between the type of classifiers and features. Two main types of features are identified: prosody contours, "short-term features", and statistics of prosody contours, "long-term features". While the former keep information on temporal dynamics, the latter loose it completely. For the classification of "short-term features" we rely on artificial neural networks (ANN), multichannel hidden Markov models, and mixture of hidden Markov models. For the classification of "long-term features", Ververidis and Kotropoulos (2006) considers two types of classifiers: those that work based on the probability density estimation (pdf) of features, and those that do not use any distribution model. Bayes classifier is an example of a classifier that relies in probability density estimation. It works on distributions modelled either by Gaussians, mixtures of Gaussians, or Parzen windows. Classifiers that do not depend on that modelling are k-nearest neighbours (kNN), support vector machines (SVM), and also artificial neural networks (ANN).

On Table 2.1 establishes connection between classifiers and works that used them.

Table 2.1: Some work references to main classifiers.

Classifier	**Work**
HMM	Fernandez and Picard (2003), Zhou et al. (2001), Hansen (1996), Hansen et al. (2011)
ANN	Fernandez and Picard (2003), Li and Zhao (1998)
SVM	Fernandez and Picard (2003), Kharat and Dudul (2008), Batliner et al. (2006), Schuller et al. (2008), Yang et al. (2012)
GMM	Lu et al. (2012), Ververidis and Kotropoulos (2005), Li and Zhao (1998)

2.8 Summary

This chapter introduced the most important concepts, as well as the most relevant works in this field. Most of the works that shall be mentioned henceforth have already been presented in this section.

Chapter 3

Searching for Stress Discriminating Feature Sets

3.1 Problem Definition

The goal of this work is to find a set of features that enables stressed speech discrimination in VOCE dataset. The stressed speech discrimination algorithm shall be feasible to run in smartphones real-time. Therefore, the feature set should have no more than a few hundred features, and the classification algorithm also be computationally light.

Previous work on the same dataset did not provide clear information on the sets that provide the best classification performance, considering most sets of features proposed in the literature. An exception to this are Teager Energy Operator-based features that are only now being applied in this dataset.

The recordings were made at a 44.1kHz (henceforth often simplified as “44kHz”) sampling frequency. The analysis of features extracted from audios at 8kHz provides a more realistic approach, however, since this is the sampling frequency used in mobile phones. It also provides noise reduction, since higher bands of spectrum, that do not bring relevant information within the approach this thesis follows, are not considered. Therefore, these audios were re-sampled to 8kHz. Thus, this thesis also explores the impact of the sampling frequency on the discrimination of stress in speech by applying the same methodology to the original dataset and one obtained for the VOCE corpus sub sampled at 8KHz.

3.2 Methodology

The methodology followed in this work comprehends sequentially the following aspects:

- extracting features;
- perform feature selection;
- evaluate and compare results.

Though not explicitly describing the methodology in this work, Figures 3.1 and 3.2 were included to illustrate the whole process. For Figure 3.1, describing training, from left to right, we have: speech signal is recorded at some moment; features are extracted; feature selection is done with a wrapper and a certain classifier; a subset of the extracted features is considered – Reduced Feature Set. For Figure 3.2, describing test, from left to right, we have: speech signal is recorded at some moment (though the speaker is never the same as in the training part); features from the previously obtained Reduced Feature Set are extracted; features of Reduced Feature Set for the training part are used to train a classifier – the same with which features were selected; the classifier classifies test utterances as stressed or neutral.

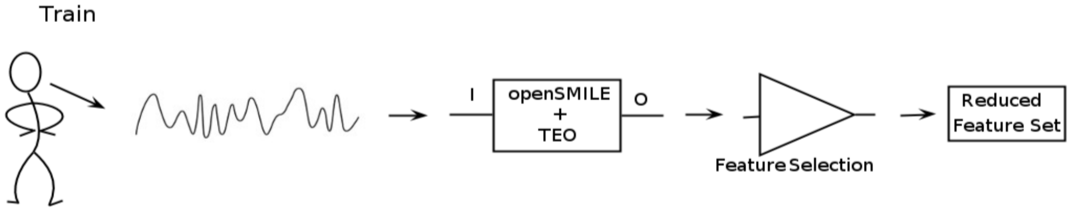

Figure 3.1: Training process

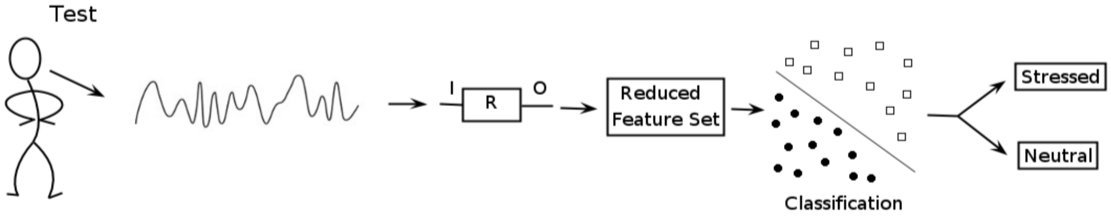

Figure 3.2: Test process

Now, concerning the scope of this work, two sets of features are firtly extracted: Feature Set I and Feature Set II. All features resulting from feature selection in this work belong either to Feature Set I ou Feature Set II. Feature Set I is automatically extracted with the openSMILE toolkit from each of the recordings. Feature Set II is automatically extracted with code developed as part of the work, except for the specific segmentation it requires, into phones, that was kindly provided by a partner project. Feature Set II is made of features based on Teager Energy Operator, that are computed for short voiced frames. Therefore, for each phone, there are as many values of a feature as frames in a phone.

Scheme in Figure 3.3 tries to help to explain this process. It shall be noted that scale is not kept between different segmentation levels (speech, utterance, phone), and "Frame Length" only refers to the unit Phone. utt1, ..., utt10 represent the utterances in a speech; ph1, ... ph8 represent the phones in an utterance; f1, ..., f4 represent the frames in a phone. For some features in Feature Set II the frame length is predetermined at some value; for others, it based on parameters extracted from the phone. For the latter, the length is the same within each phone. One TEO-based feature value is extracted por frame.

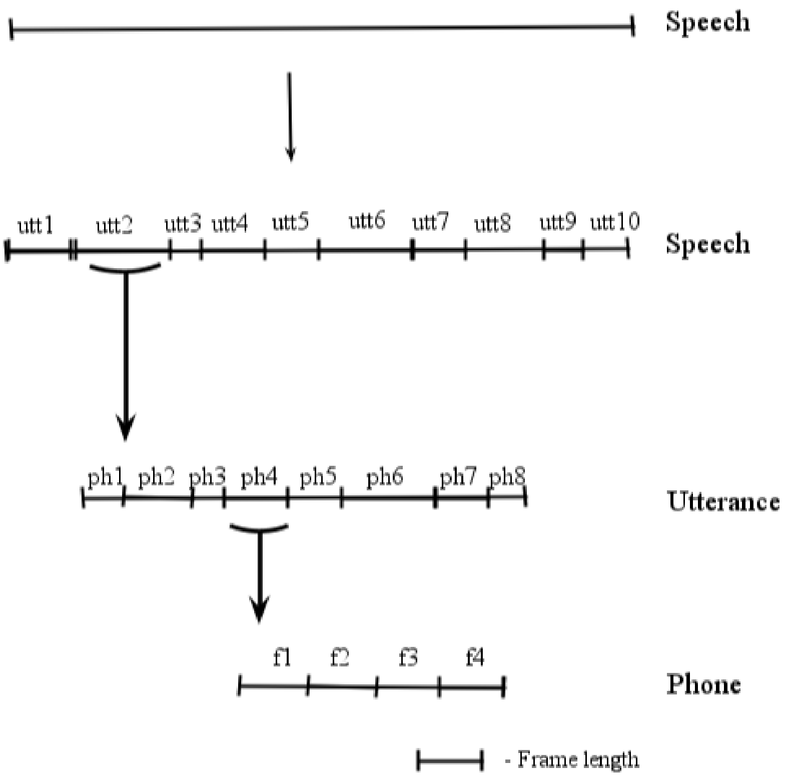

Figure 3.3: Audio segmentations.

For feature selection, the feature selection algorithm ran on three sets:

- Feature Set I;
- Informed Choice Feature Set;
- Feature Set I + Feature Set II .

Feature selection was done with a wrapper algorithm with forward selection, with 5-fold cross-validation, having embedded a SVM classifier, with C=1 and linear or gaussian kernel.

The evaluation of performance of the initial and the resulting sets was done with an SVM parametrized as the one used in feature selection, thus allowing comparison between different sets.

Feature Set combination and selection are illustrated in Figure 3.4.

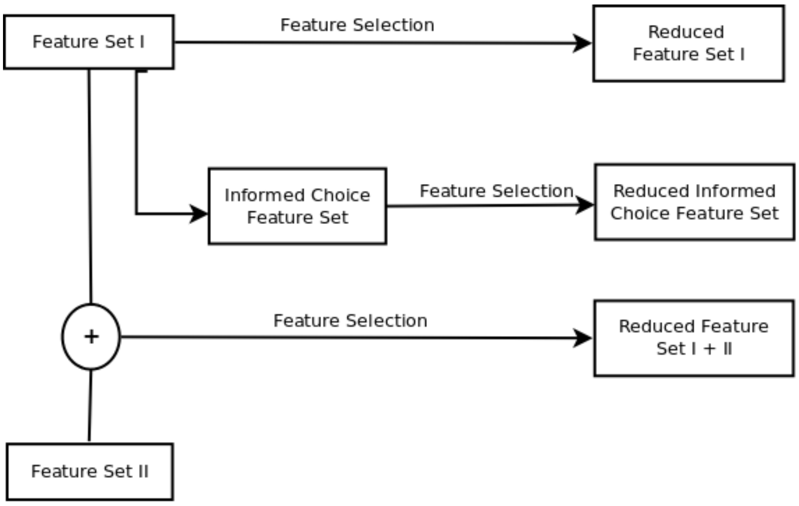

Figure 3.4: Feature Set combination and selection.

Technology

For the extraction of Feature Set I, the openSMILE toolkit has been used.

For Feature Set II, segmentation of the audios into utterances was made, in order to reduce the time of loading the audios (one at a time). This segmentation was done using ffmpeg, a program to handle multimedia data. FFmpeg was also used for re-sampling.

The computations of the second set of features were made using matlab. Voicebox - a toolbox of speech processing for Matlab - was intensively used for feature computation.

The algorithms for feature selection and for classification were build using Python. Packages Numpy and Mlpy, of scientific computing and machine learning, respectively, have been extensively used.

3.3 Data

3.3.1 VOCE Corpus

VOCE corpus, at the moment, consists of 38 raw recordings, from 28 students from University of Porto, aged 19 to 49 years old. Out of this 38 recordings, only 22 are complete, in the sense that each consists of a recording of each of the three recording moments: Baseline, Experiment, Event (prepared unscripted speech).

An ideal quality of the recording would imply, among other aspects, to make the recordings in a small room (area below 25 squared meters), or in a room with acoustic preparation, for echo and reverberation reduction. It happens that the higher the quality of the recording, the higher the complexity of the recording apparatus, and the lower the spontaneity of the speaker. Thus, there is an unavoidable trade-off between the quality of the recording and the spontaneity the speaker. It must be said that not only the quality of the recording allows more reliable data analysis, in general, but also it specifically impacts the performance of the segmentation algorithm. This is important because the utterance, as sentence-like unit, is the speech unit of the project, and the performance of the segmentation algorithm is strongly affected by noise. Therefore, due to the ecological approach followed in the project, recordings could not be done according to the procedures recommended by signal processing. And, consequently, out of the 22 complete recordings, 20 were chosen (after human listening) for having the best audio quality.

Out of the 20, one happened to be missing a Baseline sensor, which would preclude us to correctly label the utterances according to the method we are using here (v. Subsection 3.4.2). Therefore, only 19 recordings could be used in this work.

After this, 4 audios corresponding to the Event of 4 speakers did not have phone segmentation, due to low signal energy. Since this segmentation is crucial for the extraction of TEO-based features, these recordings were not considered. Consequently, only fifteen recordings of the VOCE corpus were used in this work.

3.3.2 Dataset for this work

The dataset used for this thesis was obtained from the VOCE corpus. For a matter of coherence throughout the different tasks of VOCE, train and test groups for the initially chosen 20 speakers were established. The number of utterances per speaker varies largely (v. Figure 3.5), which happens not only because speeches have different lengths, but also because the segmentation algorithm is not error free. Furthermore, 5 out of those 20 speakers have been excluded. Out of these 5 speakers, 4 were assigned to the train group, and one was assigned to the test group. The speaker that was not considered due to the lack of a Baseline sensor was assigned to the train group. Then, if these 15 speakers stood in the groups they were assigned to, train group would have 10 elements, and test group would have 5 elements. This is close to the desired splitting proportion 70% – 30%. However, this configuration would also lead to undesired proportion in the size of the sets concerning their number of utterances [1] .

Therefore, since the work unit is the utterance, and since it is desired that they are distributed approximately as 70% – 30% for train and test, respectively, speakers were relocated within these two groups. Thus, there are 4 speakers in the test set and 11 speakers in the train set (73% – 27%). Due to invalid values for some features in Feature Set II, 19 utterances were removed for 44kHz, and 18 utterances were removed for 8 kHz. These do not correspond to any speaker, speech or annotation in particular. For generalization test, testing was done on 185 utterances, for all feature

[1] This would happen mainly because the 4 speakers removed due to missing phone segmentation were the ones with higher number of utterances in the Event speech:

- For the 19 speakers, the average number of utterances per Event is 81.89 utterances;
- For 15 speakers, the average number of utterances per Event is 46 utterances.

And therefore, a huge number of utterances would be excluded from the training set. Furthermore, it is also important to notice that Event speech is the recording that contributes the most to stress annotation (v. Figures 3.6 and 3.7). As explained in the next paragraph, for the process classifier training and testing all stress utterances were taken, and the same number of neutral ones was chosen. I.e. for having long events, these speakers had many stressed utterances and, consequently many neutral utterances were considered. This is the reason why the exclusion of these speakers impacted so much the balance of train and test groups.

sets. For Feature Set I+II, feature selection was done with 370 utterances for 44kHz and 371 utterances for 8kHz. For both sampling frequencies, for Feature Set I, feature selection was done with 385 utterances. These values correspond to a distribution 67% – 33% between train and test.

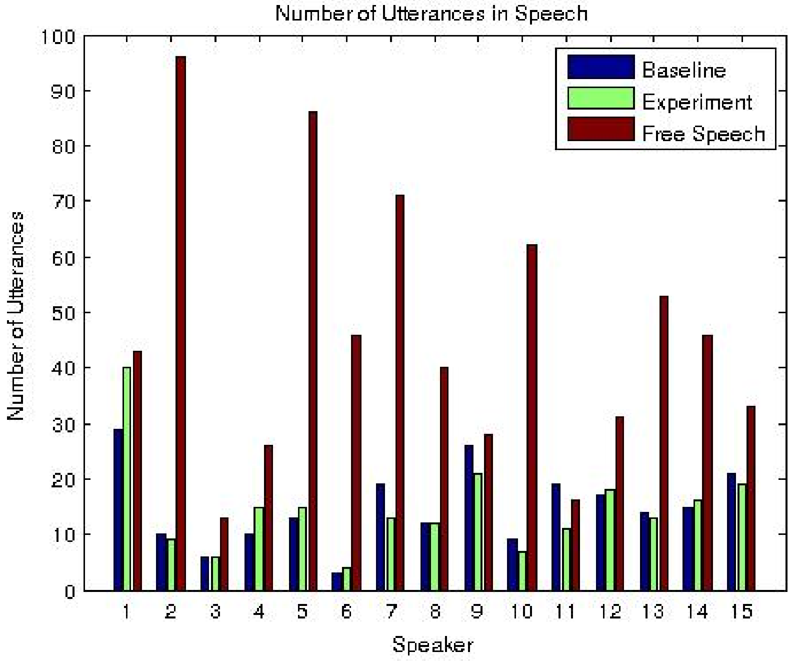

Figure 3.5: Number of utterances on the three speeches per speaker

Utterance annotation was done binarilly: each utterance was either considered as stressed or neutral (v. Subsection 3.4.2 for more details on annotation). For each speaker, the annotation was done based on the distribution of physiological annotations for all utterances of that speaker alone. Tagged utterances are the ones for which their physiological annotation is in the highest quartile of the distribution. This means that only a quarter of all utterances is labelled as stressed. Since the number of neutral utterances is three times the number of stressed utterances, there is an unbalancement problem. Therefore, using the whole set for train would bias the classifier. Due to this reason, only one third of the neutral utterances of a speaker are considered, in order to have them in the same number as stressed utterances.

Concerning the lengths of the utterances, utterances with more than 30 seconds were taken as outliers. It is suspected that those result from segmentation errors. For the set of 15 speakers, considering all available utterances after discarding outliers, the distribution of utterance lengths can be seen in the histogram of Figure 3.8.

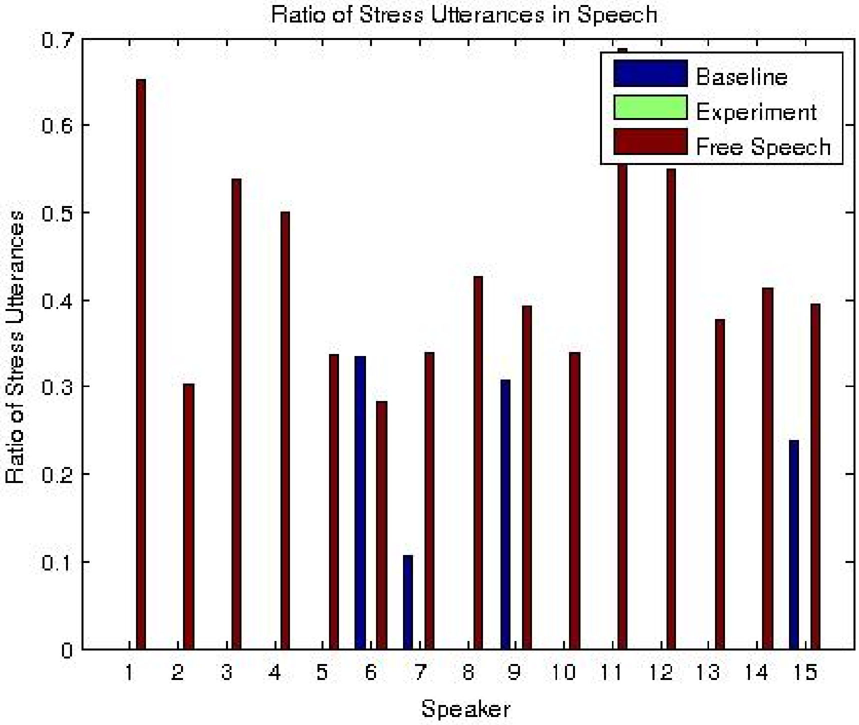

Figure 3.6: Ratio of stress-labeled utterances per speaker and speech

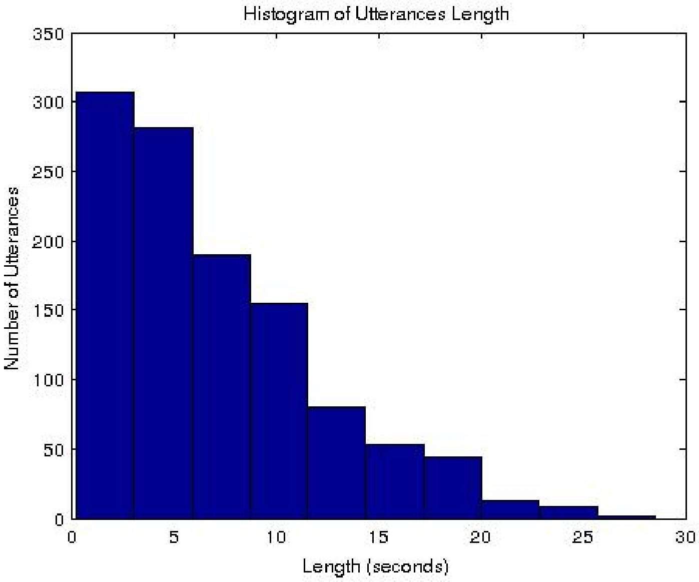

Figure 3.8: Histogram of Utterance Length

Speakers in the dataset are characterised in Table 3.1 concerning age, gender, and public speaking experience. Public speaking experience ranges from 1 to 5, where 1 stands for little and 5 stands for great. The four bottommost speakers are the ones in test set, while others are the ones

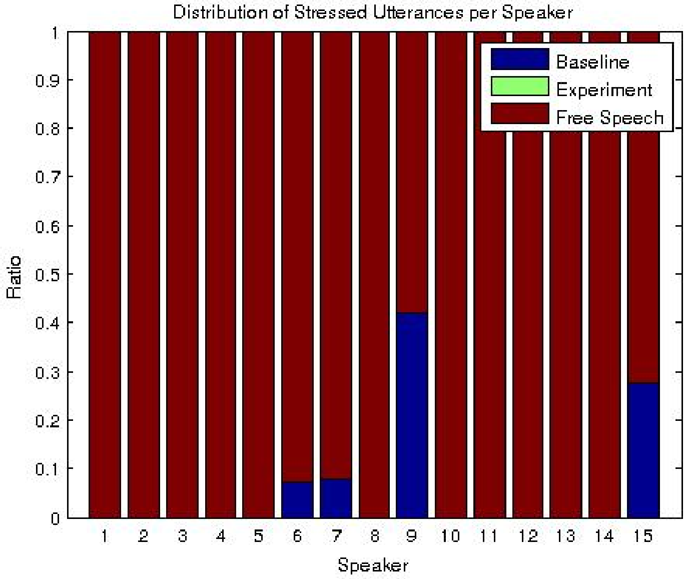

Figure 3.7: Distribution of stress-labelled utterances per speaker and speech

in training set.

Table 3.1: Dataset demographic data.

Age	**Gender**	**PS experience 1 - 5**
26	male	2
19	male	3
22	male	3
24	female	5
23	female	4
21	male	3
22	male	3
21	male	3
24	male	3
23	female	3
24	male	3
21	female	3
25	male	2
19	male	2
22	male	2

3.4 Speech Segmentation and Annotation

3.4.1 Speech Segmentation

Speeches are first automatically segmented into utterances by a speech recognition and punctuation algorithm, described in Batista et al. (2012). This algorithm reached 68.4% precision in detection of full stops for fully automatic transcripts. Since Feature Set I is extracted for whole utterances and, furthermore, since annotation is based on physiological values for whole utterances as well, the precision of the results of this work cannot be independent of the performance of this algorithm.

3.4.2 Speech Annotation

During the first part of VOCE project data analysis, stress annotation was done with the assumption that stress levels were higher in Event than in Experiment, and higher in Experiment than in Baseline. However, due to weak results this idea was abandoned, and physiological metrics based on Heart Rate (HR) were adopted instead for stress annotation. Figure 3.6 shows the ratio of stress-labelled utterances there is on each speech, for each speaker, while Figure 3.7 shows the way stress-labelled utterances are distributed per speech and speaker. Both of them show that this method for physiological annotation agree with the assumption that stress levels were higher in Event than in Experiment; nevertheless, this coexistence does not agree with the previous assumption, since for no Event all utterances were stressed. (A maximum of 0.688 was reached for Speaker 11 – v. Figure 3.6.)

Though data annotation is not the goal of this thesis, it was done as follows:

- For each speaker:
  - compute the mean heart rate in all utterances;
  - find the value corresponding to percentile 0.75 of all heart rate means;
  - classify the utterance as stressed if its mean heart rate belongs to the fourth quartile.

Chapter 4

Feature Extraction and Selection

4.1 Feature Set I

After the automatic segmentation of the speeches into utterances, some feature sets were automatically extracted. The first was the one of Low Level Descriptors – “a set of 128 frame level features extracted each 10 ms from the signal” (Ferreira and Meinedo (2013)) . The other was the one of Functional Features – “obtained by applying statistic functionals to the LLD computed over the segment. Giving us a total of 6125 segment level features.” (in Ferreira and Meinedo (2013)). The definition of these features comes in Schuller et al. (2012). There, 64 low level descriptors are provided as in Table 4.1. The 128 low-level descriptors refered in this paragraph actually correspond to an expansion of these 24 by taking first order derivatives. It was decided to only take into account the second of these feature sets. Then, the functional features are obtained by applying the statistical functionals (Schuller et al. (2012)) in Table 4.2 to the low level descriptors (64-LLD) and their first-order derivatives (128- Δ LLD). These features and their extractions processes are described in Eyben et al. (2010) and Schuller et al. (2007).

4.2 Feature Sets Selection

Feature selection was done using the wrapper method in forward selection, combined with an SVM classifier, using 5-fold utterance-based cross-validation. A scheme of this method was taken from Kohavi and John (1997), and is in Figure 4.1.

Table 4.1: Sixty-four provided low-level descriptors

4 energy related LLD
Sum of auditory spectrum (loudness)
Sum of RASTA-style filtered auditory spectrum
RMS Energy
Zero-Crossing Rate
54 spectral LLD
RASTA-style auditory spectrum, bands 1-26 (0-8kHz)
MFCC 1-14
Spectral energy 250-650 Hz, 1 k-4kHz
Spectral Roll Off Point 0.25, 0.50, 0.75, 0.90
Spectral Fluz, Entropy, Variance, Skewness, Kurtosis, Slope, Psychoacoustic Sharpness, Harmonicity
6 voicing related LLD
F0 by SHS + Viterbi smoothing, Probability of voicing logarithmic HNR, Jitter (local, delta), Shimmer (local)

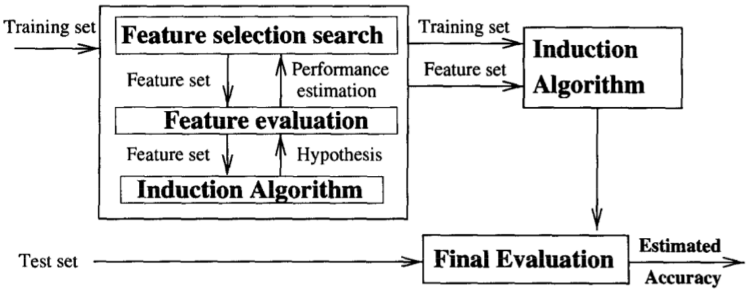

Figure 4.1: Wrapper for Feature Subset Selection

This method progressively appends to a feature set the features that provide an increment to the classification performance. This happens until none of the available features can increase the performance of the existing set. It works by running cycle on the set of features not yet chosen such that for each feature it (1) appends it to the existing set of chosen features, (2) trains and tests a classifier with that set, (3) stores the classification error. At the end of this cycle, if any of these auxiliary sets performed better than the preexisting one, the tested feature is appended to the main set (the preexisting one), and the cycle runs once more, to search for another eventual feature that helps to improve the performance of the classifier. If none of these auxiliary sets performed better than the preexisting one, the algorithm stops and returns the best set and its classification error. The pseudo-code for this algorithm can be found in Subsection 4.2.1.

Applying 5-fold cross-validation means that 5 combinations of train and test set are considered.

The chosen features are always the ones to which the minimum average classification error over those 5 combinations occured.

This algorithm is good in avoiding redundancy, it is not good, however, at finding an optimal

Table 4.2: Applied functionals: arithmetic mean of LLD / positive Δ LLD (1); only applied to voice related LLD (2); not applied to voice related LLD except F0 (3); only applied to F0 (4).

Functionals applied to LLD / Δ LLD
quartiles 1–3, 3 inter-quartile ranges
1 % percentile (≈ min), 99 % percentile (≈ max)
position of min / max
percentile range 1 % – 99 %
arithmetic mean (1), root quadratic mean
contour centroid, flatness
standard deviation, skewness, kurtosis
rel. duration LLD is above / beloow 25/50/75/90 % range
rel. duration LLD ir rising / falling
rel. duration LLD has positive / negative curvature (2)
gain of linear prediction (LP), LP Coefficients 1–5
mean, max, min, std. dev. of segment length (3)
Functionals applied to LLD only
mean of peak distances
standard deviation of peak distances
mean value of peaks
mean value of peaks – arithmetic mean
mean / std.dev. of rising/falling slopes
mean / std.dev. of inter maxima distances
amplitude mean of maxima/minima
amplitude range of maxima
linear regression slope, offset, quadratic error
quadratic regression a, b, offset, quadratic error
percentage of non-zero frames (4)

set. Since it stops at the point where no feature else can improve the performance, set resulting from this classifier tend to be very small.

4.2.1 Implementation

The implementation of the previous algorithm is described bellow.

```
set empty vector of features – I;
set empty vector of errors – I;

set empty vector of features – II;
set empty vector of errors – II;


for each feature – feature0:
        train classifier with feature0 only;
```

```
    test classifier with feature0 only - error0;

    append feature0 to vector II;
    append error0 to vector II;


choose feature to which corresponds the smallest
classification error - feature1, error1;
append feature1 to vector of features I;
append error1 to vector of errors I;


while True:

    set vector of features II equal to vector of features I;
    set empty vector of errors - II;

        for each feature not in vector of features I
         - featureT:

            append featureT to vector of features II;
            train the classifier with vector of features II;
            test classifier with that feature only - errorT;

            append errorT to vector II;

        choose feature to which corresponds the smallest
        classification error: featureC, errorC;

        if error2 >= last error in vector of errors I:
            stop algorithm
        else:
            append featureC to vector of features I;
            append errorC to vector of errors I;

return vector of features I;
return last error.
```

4.3 Informed Choice Feature Set

Among all the popular features in speech processing, some can *a priori* be said to be more likely to provide good results in stress classification, which is an alternative to lighten the load for the classifiers. For instance, Demenko and Jastrzebska (2012) mention Pitch Variation, Noise to Harmonics Ratio, Subharmonics, and Voice Irregularities. Lu et al. (2012) talk about Pitch (mean, standard deviation, range), Pitch Jitter, and TEO-based features. In this case, Lu et al. (2012), concerning their own achievements MFCC is the most relevant base for features indoor (TEO is outdoor). We did not find in the literature large deviations from this set. As was mentioned in Chapter 2, pitch and features related to voice quality are the most frequently mentioned ones. VOCE consultors proposed as good features for stress discrimination Jitter, Shimmer, Harmonics to Noise Ratio and Subharmonics to Harmonics Ratio.

The Informed Choice feature set was obtained as a subset of Feature Set I. Subharmonics to Harmonics Ratio were not considered for not belonging to Feature Set I. Voicing features were considered as measures of voice quality.

We clearly exclude here TEO-based features, in order to include them in the next sets. We have, therefore a set of 460 features that is discriminated in Table 4.3.

Table 4.3: Sets of long-term features as Informed Choice

Informed Choice Features
voicingFinalUnclipped
jitterLocal
jitterDDP
shimmerLocal
logHNR

4.4 Feature Set II

Feature Set II is composed by TEO-Based features extracted in the context of this work. Except for Variation of FM component, we tried to reproduce the following TEO-Based features: Normalized TEO autocorrelation envelope, Critical band based TEO autocorrelation envelope, and TEO-based cepstal coefficients. Although for TEO-based cepstal coefficients the proposed implementation is easy to follow, the same does not happen with the other features. We contacted the authors of the Zhou et al. (2001), in order to have details on the way to compute features and we were given very limited information due to sponsoring policies. It must be said, then, that since these algorithms were implemented based only on the articles, some details were left to our choice or implemented differently (as an attempt to find easier solution). More information on this can be found in Subsection 4.4.6.

This section starts by describing how data was parsed prior to feature extraction in Subsection 4.4.1; then the statistical treatment of features is presented in Subsection 4.4.2. Feature extraction processes are described in Subsections 4.4.3, 4.4.4, and 4.4.5. Finally implementation details are stated in Subsection 4.4.6.

4.4.1 Phone Selection

The literature where Normalized TEO autocorrelation envelope are Critical band based TEO autocorrelation are presented does the feature extraction for small voiced parts usually called "tokens". These are, for instance, "freeze", "help", "mark", "nav", "oh", and "zero" (Zhou et al. (2001)). To work equivalently, we were kindly provided by a project partner a phone transcription of the texts, with the delimitation of each phone. Only vowels were chosen, since these features are computed upon voiced speech, and due to the "definite quasi-periodic nature" of vowels (Ruzanski et al. (2005)). Lu et al. (2012) justifies that TEO-CB-AutoEnv features, as well as pitch, only from voiced frames can be properly extracted, and also that since voice speech is significantly more energetic than invoiced speech, it is more resilient to ambient noise. The phones used here are the ones corresponding to portuguese SAMPA symbols 'i','e','E','a','6','o','O','u','@'. (SAMPA - Speech Assessment Methods Phonetic Alphabet) [1].

4.4.2 Statistics

These features are extracted per frame. Each phone usually contains many frames. Each utterance has normally many phones. Therefore, since we want to have values per utterance, we consider each feature extracted for all phones and apply statistics to it. This process is illustrated in Subsection 4.4.6.

These statistics are: mean, standard deviation, skewness, kurtosis, first quartile, median, third quartile, and inter-quartile range.

4.4.3 Normalized TEO Autocorrelation Envelope

This feature was made to "capture stress dependent information that may be present in changes within the FM component" Zhou et al. (2001). It comes from the model proposed by Maragos et al. (1993) where "voiced speech can be modeled as the sum of AM-FM signals of which each is centered at a formant frequency". Given this, a filter bank can be used to bandpass filter voiced speech around each of its formant frequencies, being the modulation pattern across each of the formants obtained using TEO. As it is not feasible to track all formant frequancies to bandpass filter based on them, the spectrum is divided into four bands: 0-1 kHz, 1-2 kHz, 2-3 kHz, 3-4 kHz. It is expected that under stress formant frequencies may possibly shift to other frequency bands.

The TEO profile of each band is filtered arounf F0, with bandwidth of F0/2 in order to capture variations around F0. Then, in order to average effects for the formant variations each part is segmented in frames with 4 median pitch periods of length.

For each frame, the normalized autocorrelation function is computed. According to the authors, "if there is no pitch variation within a frame, the output TEO is a constant". Then the area

[1] http://www.phon.ucl.ac.uk/home/sampa/portug.htm

under the normalized autocorrelation envelope is computed and normalized by N/2, where N is the number of samples per frame. The obtained metrics reflect the degree of excitation variability within each band.

So, the feature is computed as follows:

```
compute F0;

filter the signal in 4 bandpass filters with
bandwidths 0-1 kHz, 1-2 kHz, 2-3 kHz, 3-4 kHz,
and centres 0.5 kHz, 1.5 kHz, 2.5 kHz, 3.5 kHz, respectively;

for each resulting band:
        compute TEO;

for each TEO band:
        bandpassfilter with centre F0 and bandwidth F0/2;

for each filtered TEO band:
        segment into frames of 4 median pitch periods;

for each segmented filtered TEO-band:
        for each frame in  segmented filtered TEO-band:
                compute autocorrelation;

for each segmented filtered TEO-band:
        for each frame in  segmented filtered TEO-band:
                compute area under the autocorrelation function;
                normalize the area under autocorrelation
                by the length of the frames.
```

This procedure is illustrated in Figure 4.2.

4.4.4 Critical Band Based TEO Autocorrelation Envelope

Critical Band Based TEO Autocorrelation Envelope relies on the same principles as normalized TEO autocorrelation envelope, but is done with a finer partition of the spectrum. Here the considered frequency bands are the ones in Bark scale until 3700 Hz – v. Table 4.4 . This scale models the human auditory system by mapping frequencies into critical bands, where the responses to stimuli in the ear are similar (Smith and Abel (1999)).

It is computed as follows.

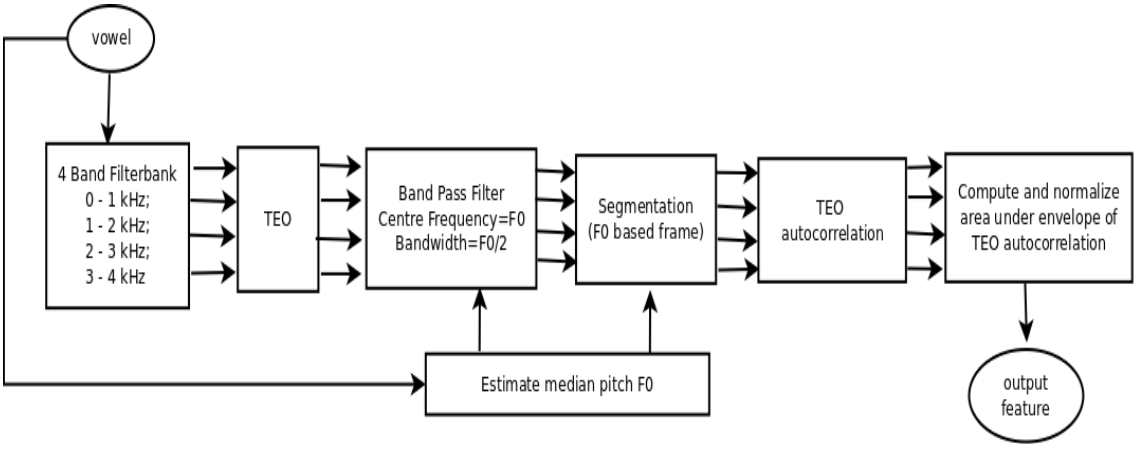

Figure 4.2: Normalized TEO Autocorrelation Envelope Extraction Process

```
filter the signal with a filterbank into Bark critical bands;

for each frequency band filtered signal:
    compute TEO;

for each filtered TEO profile:
    segmented into frames of 25 ms with 12.5 ms of overlap;

for each segmented filtered TEO profile:
    for each frame in the segmented filtered TEO profile:
        compute autocorrelation;

for each segmented filtered TEO profile:
    for each frame in the segmented filtered TEO profile:
        compute area under autocorrelation envelope;
        normalize area under autocorrelation envelope by
        half of the number of samples in a frame .
```

This procedure is illustrated in Figure 4.3. This feature has the advantage of not being dependent on F0 estimation.

4.4.5 TEO-based “cepstral coefficients”

In Jabloun et al. (1999) TEO-based “cepstral coefficients” due to the experimental knowledge that TEO did suppress car engine noise, it was supposed that it could be useful to recognition systems. The authors come from the following reasoning: recorded signal (x(t)) is the sum of speech signal (s(t)) and noise (v(t)) – zero-mean and independent from s(t).

Table 4.4: Critical band frequency information: bark scale

	Critical band frequency information (Hz)			
Band number	Lower	Centre	Upper	Bandwidth
1	100	150	200	100
2	200	250	300	100
3	300	350	400	100
4	400	450	510	110
5	510	570	630	120
6	630	700	770	140
7	770	840	920	150
8	920	1000	1080	160
9	1080	1170	1270	190
10	1270	1370	1480	210
11	1480	1600	1720	240
12	1720	1850	2000	280
13	2000	2150	2320	320
14	2320	2500	2700	380
15	2700	2900	3150	450
16	3150	3400	3700	550

This way, the Teager energy of the signal is

$$\Psi[x(n)] = \Psi[s(n)] + \Psi[v(n)] + 2\Psi[s(n)v(n)] \quad , \tag{4.1}$$

where

$$\Psi[s(n)v(n)] = s(n)v(n) - \frac{1}{2}s(n+1)v(n-1) - \frac{1}{2}s(n-1)v(n+1) \tag{4.2}$$

is the cross energy of s(n) and v(n). Due to the fact that s(n) and v(n) are independent and v(n) is zero-mean, this term is null. Therefore, we get

$$\Psi[x(n)] = \Psi[s(n)] + \Psi[v(n)] \quad . \tag{4.3}$$

Since for car noise $\Psi[s(n)] >> \Psi[v(n)]$, it is taken that

$$\Psi[x(n)] \approx \Psi[s(n)]. \tag{4.4}$$

This shows TEO's noise filtering capacity, unlike the commonly used energy, $E[x^2(n)]$.

To obtain the new features, the speech signal in spectrally divided into 21 non-uniform subbands in mel-scale, shown in Table 4.5. For each of these subbands, the Teager energy is estimated and a feature vector is build by compression and inverse discrete cosine transform. Approach of Fernandez and Picard (2003) in which for compression whole frames are used is taken here.

In practice, after the decomposition in frequency bands the energy coefficients are computed,

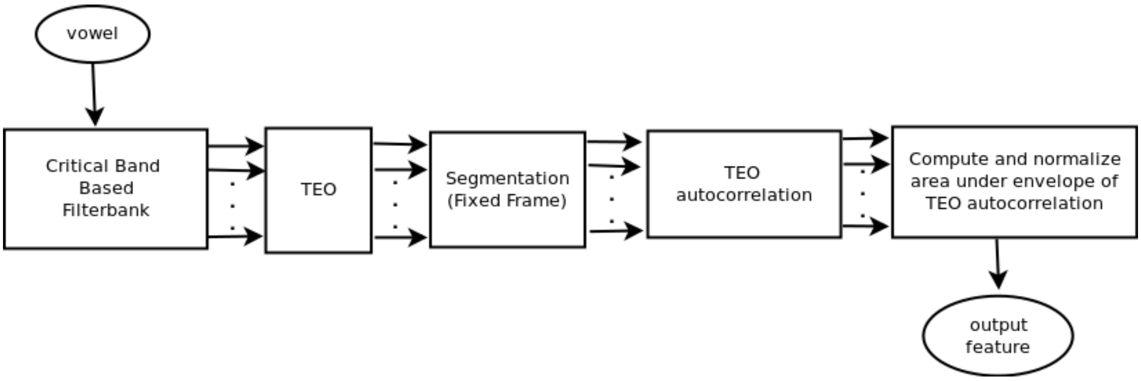
vowel
Critical Band Based Filterbank
TEO
Segmentation (Fixed Frame)
TEO autocorrelation
Compute and normalize area under envelope of TEO autocorrelation
output feature

Figure 4.3: Critical Band Based TEO Autocorrelation Envelope Extraction Process

for each frequency band, with

$$e_m = \frac{1}{N_m}\sum_{n=1}^{N_m} |\Psi[x(n)]| \quad ,m = 1,...,M = 21 \quad , \tag{4.5}$$

being N_m the number of time samples in the m^{th} band and M the number of frequency subbands (21).

The TEO-based “cepstrum coefficients” are then obtained through the application of an inverse DCT transform to the log of the energy coefficients:

$$E_l = \sum_{m=1}^{M} log(e_m) cos\left(\frac{l(m-0.5)\pi}{M}\right) \quad ,l = 1,...,L \tag{4.6}$$

being L the number of bands to which we want to compress the original M. Here L is set to 10. The frames have 24 ms of length and 10 ms of overlap between adjacent frames.

For the r^{th} frame, the vector of cepstral coefficients is defined:

$$\mathbf{E}^{[r]} = [E_1^{[r]},...,E_L^{[r]}] \quad . \tag{4.7}$$

As a metric for autocorrelation between frames, the following is proposed:

$$ACE_{l,\tau}^{[r]} = \frac{\sum_{n=r}^{r+T} E_l^{[n]} E_l^{[n+\tau]}}{max_j(\sum_{n=r}^{r+T} E_l^{[n]} E_l^{[n+\tau]})} \quad l = 1,...,L \quad , \tag{4.8}$$

where τ is the lag between frames (=1, since we want them adjacent), T is the number of frames in the autocorrelation window (=2, since we want to consider pairwise correlation), and j is an index that spans all correlation coefficients within the same scale along all frames to normalize the autocorrelation.

Then, the vector containing the logarithms of the autocorrelation coefficients is defined:

$$\mathbf{ACE\_L}_{\tau}^{[r]} = [logACE_{1,\tau}^{[r]},...,logACE_{L,\tau}^{[r]}] \quad . \tag{4.9}$$

Table 4.5: Twenty-one subband decomposition

	Subband frequency information (Hz)		
Band number	Lower	Upper	Bandwidth
1	0.	62.5	62.5
2	62.5	125.	62.5
3	125.	187.5	62.5
4	187.5	250.	62.5
5	250.	312.5	62.5
6	312.5	375.	62.5
7	375.	437.5	62.5
8	437.5	500.	62.5
9	500.	625.	125.
10	625.	750.	125.
11	750.	875.	125.
12	875.	1000.	125.
13	1000.	1125.	125.
14	1125.	1250.	125.
15	1250.	1375.	125.
16	1375.	1500.	125.
17	1500.	1750.	250.
18	1750.	2000.	250.
19	2000.	2500.	500.
20	2500.	3000.	500.
21	3000.	4000.	1000.

The final feature vector, $\mathbf{FS}^{[r]}$, comes from the concatenation of $\mathbf{E}^{[r]}$ and $\mathbf{ACE\_L}_{\tau}^{[r]}$,

$$\mathbf{FS}^{[r]} = [\mathbf{E}^{[r]} \quad \mathbf{ACE\_L}_{\tau}^{[r]}] \quad . \tag{4.10}$$

4.4.6 TEO-Based Features Implementation

Figure 4.4 helps to understand this section. It represents the extraction of one feature. Please notice that three groups of features compose Feature Set II, and that each group is composed of several features, corresponding to frequency bands, or equivalent representations of frequency bands. The extraction represented in Figure 4.4 addresses only one of the latter.

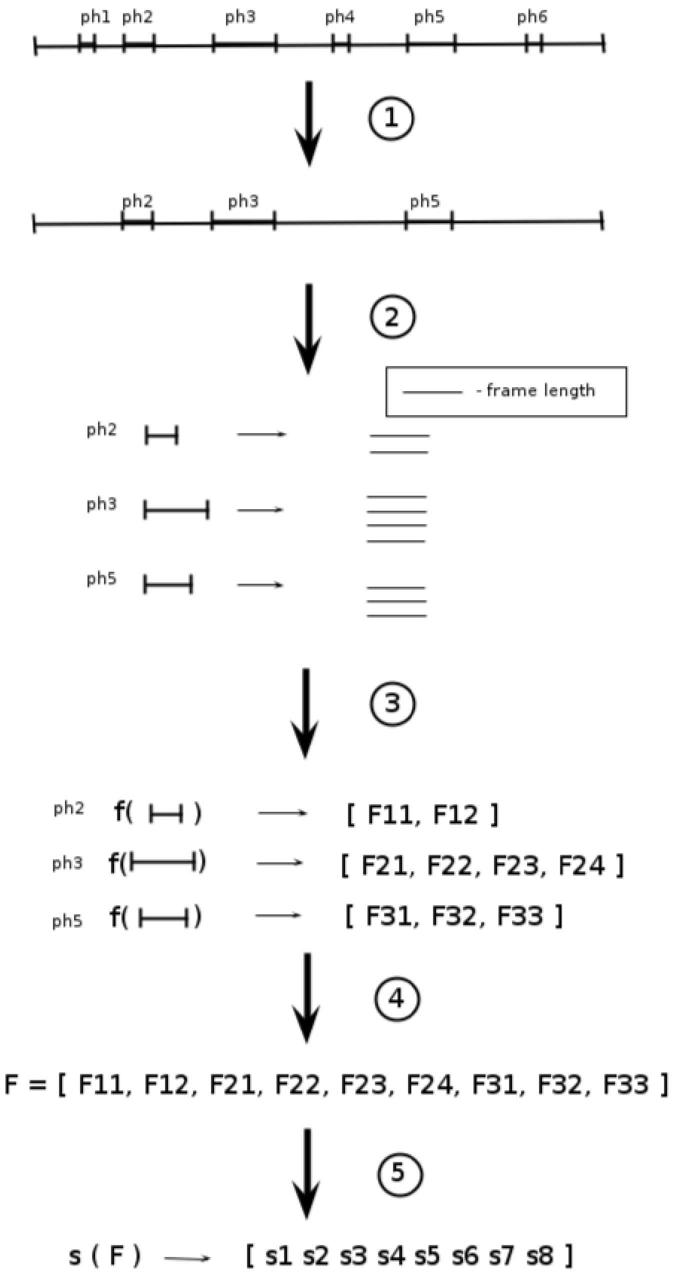
ph1 ph2 ph3 ph4 ph5 ph6
1
ph2 ph3 ph5
2
- frame length
ph2
ph3
ph5
3
ph2 f( ) → [F11, F12]
ph3 f( ) → [F21, F22, F23, F24]
ph5 f( ) → [F31, F32, F33]
4
F = [F11, F12, F21, F22, F23, F24, F31, F32, F33]
5
s (F) → [s1 s2 s3 s4 s5 s6 s7 s8]

Figure 4.4: Extraction process of a feature from Feature Set II.

The steps of the process are described below.

1. The longest line represents the utterance and the short delimited and labelled lines ont it represent the phones (ph1,...,ph5). This part is phone selection, where only voiced phones, with more than 50 ms are selected. Selection also used Voice Activity Detection (VAD) and Voiced-Unvoiced Detection (VUD) [2] : only phones with at least 85% of their frames classified as having voice activity and with all the frames having voice detected were considered.

[2] VAD does detection of human voice, while VUVD identifies voiced parts, i.e., excludes unvoiced consonants.

For the diagram in Figure 4.4, phones ph1, ph4, and ph6 were discarded. For Voice Activity Detection (VAD) and Voiced-Unvoiced Detection (VUD) we used functions *vadsohn* and *fxrapt* of voicebox, of matlab.

2. Selected phones are divided into frames (with or without overlap, depending on the feature) of fixed length using function *enframe* from voicebox.
3. Features are computed for each phone, returning as many values as frames in the utterance.
4. All features are concatenated into a vector, F.
5. Statistics are applied to the vector F, returning vector s as output. The entries of s are the final features.

The exposed procedures, though corresponding to an ideal way of implementation, were not feasible. Errors came from many sources, and it was not possible to identify them all. The most understandable ones had to do with length: for Normalized TEO autocorrelation envelope, as frame length is based on median pitch period, frame length varies a lot. Due to this fact, imposing minimal lengths to phones may cause an avoidable and undesired waste of data. Other errors came from bandpass filters, but no further details were found. In practice, and to obtain results, functions for feature computation were applied to all vowel phones. Phones that led to errors after applying feature functions were descarded.

It also happened that even when features led to results, these were not numbers (NaN). As the classifier did not run for data with NaN, (1) features whose values were all NaN were removed, and after, (2) vectors with at least one value NaN were also removed. In practice this was feature selection too, since 80 features were removed, both for extractions at 44.1kHz and at 8kHz. These correspond to TEO-based “cepstral coefficients”. From the set at 44.1kHz, 19 vectors were removed, and from the set at 8kHz, 18 vectors were removed. No particular incidence on speaker, class or speech event was noticed in the removed sets.

As said in the beginning of this section, 4.4, some implementation details were left to our choice, and some were made differently from described for technical reasons. These are:

- the median pitch estimation is done computing the inverse of the median frequency estimation; (1)
- instead of using a gabor filter to select the frequency bands, we used a gammatone filter (bank), provided by matlab; (1)(2)(3)
- the area under the TEO autocorrelation envelope was computed using matlab function *trapz*, that computes the aproximate integral of its argument using the trapezoidal method with unit spacing. (1)(2)

(1) applied to Normalized TEO autocorrelation envelope; (2) applied to Critical band based TEO autocorrelation envelope; (3) applied to TEO-based “cepstral coefficients” .

Chapter 5

Results and Discussion

5.1 Results

The selected feature sets can be found in Tables 5.1 and 5.2 - for 44kHz, and in Tables 5.3 and 5.4 - for 8kHz. Train classification rate corresponds to the ratio of correctly classified utterances in the feature selection process; Test classification rate corresponds to the ratio of correctly classified utterances in the Test Group, for generalisation. At the end, Tables 5.5 and Table 5.6 summarize the results for 44kHz and 8kHz, respectively.

Some of the results in Tables 5.5 and 5.6 are plotted in Figures 5.1 and 5.2, for comparison of the performance of the linear and gaussian SVM kernels with the various selected feature sets on train set and test set. Figure 5.1 corresponds to the results for 44kHz, while Figure 5.2 corresponds to the results for 8kHz.

Table 5.1: Selected features, 44kHz, linear kernel.

Feature Set I	**Informed Choice**	**Feature Set I+II**
- F0final_ sma_ lpc0 - audSpec_ Rfilt_ sma[2]_ quartile1 - audSpec_ Rfilt_ sma[24]_ iqr2-3 - pcm_ zcr_ sma_ de_ iqr1-3 - audSpec_ Rfilt_ sma[4]_ downleveltime75	- jitterDDP_ sma_ iqr1-3 - jitterLocal_ sma_ percentile1.0 - voicingFinalUnclipped_ sma _ pctlrange0-1 - shimmerLocal_ sma_ quartile1 - shimmerLocal_ sma_ qregc2 - jitterLocal _sma_ range - shimmerLocal_ sma_ downleveltime90 - voicingFinalUnclipped_ sma_ lpc1 - voicingFinalUnclipped_ sma_ iqr1-3	- F0final_ sma_ lpc0 - Critical Band Based TEO autocorrelation envelope_ band16_ median - jitterLocal_ sma_ de_ rightctime - audSpec_ Rfilt_ sma[4]_ peakMeanMeanDist - mfcc_ sma[1]_ upleveltime50 - pcm_ Mag_ psySharpness_ sma_ linregc1 - audSpec_ Rfilt_ sma_ de[16]_ range - pcm_ RMSenergy_ sma_ upleveltime90 - pcm_ Mag_ spectralRollOff25.0_ sma_ de_ quartile3
Train Recognition Rate: 0.787	Train Recognition Rate: 0.779	Train Recognition Rate: 0.846
Test Recognition Rate: 0.676	Test Recognition Rate: 0.676	Test Recognition Rate: 0.741

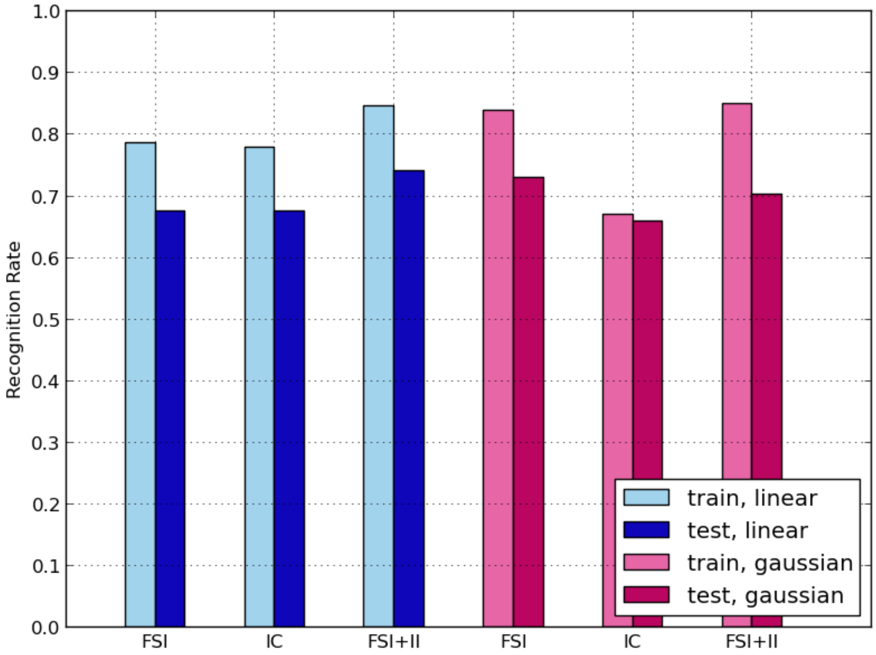

Figure 5.1: Recognition Rates for Sets Selected at 44kHz. FSI: Feature Set I; IC: Informed Choice Feature Set; FSI+II: Feature Set I+II.

Table 5.2: Selected features, 44kHz, gaussian kernel.

Feature Set I	**Informed Choice**	**Feature Set I+II**
- audSpec_ Rfilt_ sma[10]_ quartile3 - shimmerLocal_ sma_ meanPeakDist - audspecRasta_ lengthL1norm_ sma_ de_ iqr2-3 - audSpec_ Rfilt_ sma_ de[23]_ pctlrange0-1 - audspecRasta_ lengthL1norm_ sma_ qregc1 - audSpec_ Rfilt_ sma[7]_ upleveltime50 - audSpec_ Rfilt_ sma_ de[1]_ lpc4	- voicingFinalUnclipped_ sma_ range	- audSpec_ Rfilt_ sma[10]_ quartile3 - audSpec_ Rfilt_ sma[16]_ lpc3 - pcm_ Mag_ spectral-RollOff25.0_ sma_ range - audSpec_ Rfilt_ sma[14]_ downleveltime90 - audSpec_ Rfilt_ sma[7]_ quartile2 - F0final_ sma_ rqmean - pcm_ Mag_ psySharpness_ sma_ de_ lpc0
Train Recognition Rate: 0.839	Train Recognition Rate: 0.670	Train Recognition Rate: 0.849
Test Recognition Rate: 0.730	Test Recognition Rate: 0.659	Test Recognition Rate: 0.703

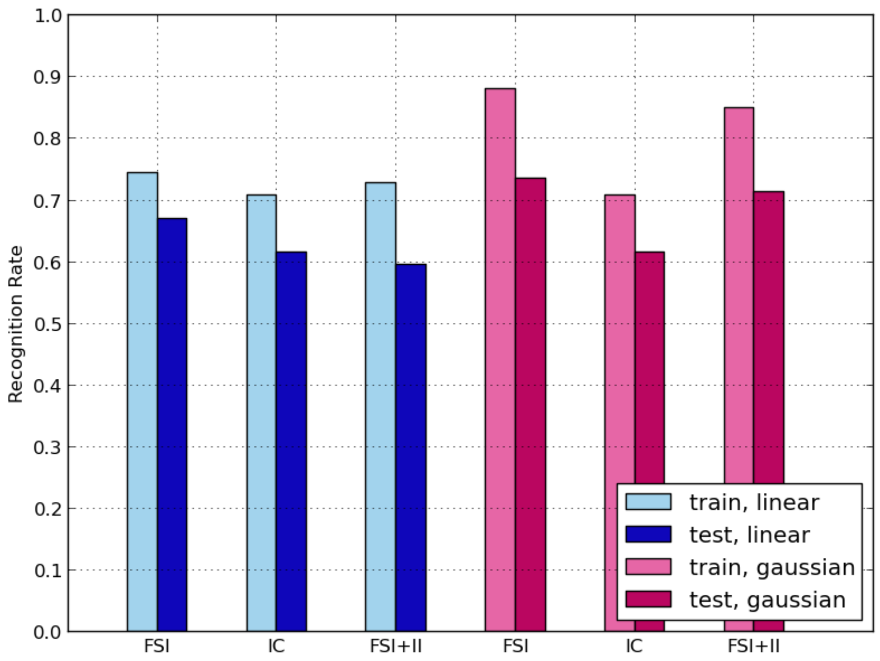

Figure 5.2: Recognition Rates for Sets Selected at 8kHz. FSI: Feature Set I; IC: Informed Choice Feature Set; FSI+II: Feature Set I+II.

Table 5.3: Selected features, 8kHz, linear kernel.

Feature Set I	**Informed Choice**	**Feature Set I+II**
- jitterLocal_ sma_ quartile1 - pcm_ Mag_ spectralRollOff50.0_ sma_ de_ quartile2 - jitterDDP_ sma_ amean - audSpec_ Rfilt_ sma_ de[20]_ lpgain - pcm_ Mag_ spectralRollOff75.0_ sma_ percentile1.0 - pcm_ RMSenergy_ sma_ de_ lpgain - pcm_ Mag_ fband1000-4000_ sma_ quartile1 - pcm_ Mag_ fband1000-4000_ sma_ percentile1.0	- jitterLocal_ sma_ quartile1 - shimmerLocal_ sma_ de_ lpgain	- jitterLocal_ sma_ quartile1 - jitterLocal_ sma_ de_ quartile1 - pcm_ Mag_ spectralRollOff75.0_ sma_ percentile1.0 - pcm_ zcr_ sma_ de_ quartile3 - pcm_ RMSenergy_ sma_ percentile1.0
Train Recognition Rate: 0.745	Train Recognition Rate: 0.709	Train Recognition Rate: 0.728
Test Recognition Rate: 0.670	Test Recognition Rate: 0.616	Test Recognition Rate: 0.595

The chosen features belong to 14 different categories. Figures 5.3 and 5.4 show the total frequency of those categories for Feature Set I and Feature Set I+II together, chosen using both linear and gaussian SVM kernels. Features chosen for the Informed Choice Feature Set were not considered for this analysis, for that is a very reduced feature space, and its comparison to others concerning feature choice would not be relevant – it would lead to biases in the histogram; features from the Informed Choice Feature Set would be more noted. Features chosen from Feature Set I and Feature Set I+II that also belong to the Informed Choice Feature Set correspond to the darker bars, in the left-hand side of the plot. Figure 5.1 corresponds to the results for 44kHz, while Figure 5.2 corresponds to the results for 8kHz.

For 44kHz, 12 out of the 38 chosen features belong to the Informed Choice Feature Set; for 8kHz, 8 out of the 35 chosen features belong to the Informed Choice Feature Set. These correspond to 33.2% and 22.9%, respectively. This is an indicator that the feature selection is returning expected results, by returning features that are expected to be correlated with stress.

Table 5.4: Selected features, 8kHz, gaussian kernel.

Feature Set I	**Informed Choice**	**Feature Set I+II**
- audSpec_ Rfilt_ sma[21]_ percentile1.0 - audSpec_ Rfilt_ sma[2]_ quartile1 - mfcc_ sma[5]_ lpc2 - pcm_ Mag_ harmonicity_ sma_ quartile3 - pcm_ Mag_ fband250-650_ sma_ rqmean - audSpec_ Rfilt_ sma_ de[20]_ iqr1-3 - audspec_ lengthL1norm_ sma_ peakMeanAbs - audSpec_ Rfilt_ sma[14]_ peakDistStddev - pcm_ RMSenergy_ sma_ upleveltime75 - pcm_ RMSenergy_ sma_ meanSegLen - pcm_ Mag_ harmonicity_ sma_ de_ quartile1	- jitterLocal_ sma_ quartile1 - shimmerLocal_ sma_ de_ lpgain	- audSpec_ Rfilt_ sma[21]_ percentile1.0 - audSpec_ Rfilt_ sma[2]_ quartile1 - mfcc_ sma_ de[5]_ quartile2 - mfcc_ sma_ de[2]_ flatness - pcm_ Mag_ spectralVariance_ sma_ de_ flatness - audSpec_ Rfilt_ sma_ de[22]_ upleveltime25 - pcm_ zcr_ sma_ de_ upleveltime25
Train Recognition Rate: 0.881	Train Recognition Rate: 0.709	Train Recognition Rate: 0.849
Test Recognition Rate: 0.735	Test Recognition Rate: 0.616	Test Recognition Rate: 0.714

5.2 Discussion

Among the obtained feature sets, none has shown an outstanding performance. However, the best recognition rate was 0.741 for Reduced Feature Set I+II, for audios sampled at 44kHz and for an SVM with linear kernel, with 9 features chosen. Some of the results presented in the literature for stressed speech discrimination are presented in Table 5.7. As stated in Vogt et al. (2008), it is rarely possible to compare features across published work, due to all the differences in implementations. Nevertheless, these values indicate that the combination of classifier and chosen feature sets achieve similar precision to what is common in similar studies.

The difference between training and test recognition rate has a maximum of 0.146 and a minimum of 0.011. Considering all sets, the difference between train and test recognition rate had mean 0.106 and variance 0.001.

Considering all feature sets, the generalisation decreased the precision in 0.106 in average, which shows robustness for the combination of feature sets and classifier. Though it is not frequent to reveal this information in literature, Fernandez and Picard (2003) does it. For a subject-

Table 5.5: Performance of the Selected Feature Sets, sampling at 44kHz. SF is the Sampling Frequency at which audio files were recorded. Train. RR is the Recognition Rate obtained in the feature selection. Gen. RR is the Recognition Rate in the test set. CS. RR is the Recognition Rate when training and testing on the full feature set.

Feature Set	**kernel**	**# Features**	**Train. RR**	**Gen. RR**	**CS RR**
Feature Set I	linear	5	0.787	0.676	0.0
Informed Choice Feature Set	linear	9	0.779	0.676	0.604
Feature Set I + II	linear	9	0.846	0.741	0.0
Feature Set I	gaussian	7	0.839	0.730	0.492
Informed Choice Feature Set	gaussian	1	0.67	0.659	0.492
Feature Set I + II	gaussian	7	0.849	0.703	0.492

dependent task, SVM classifier achieves as mean recognition rate 0.595 in training and 0.467 in test. This represents a decrease of precision of 0.128, which is greater than the decrease obtained here. Furthermore, this result is for a speaker-dependent task, while all results concerning the work presented here are speaker-independent.

Concerning the selected types of features, in Figures 5.3 and 5.4, audSpec_Rfilt and pcm_Mag are the most frequently chosen features. Features from Informed Choice Feature Set were frequently chosen – corresponding to 33.2% of the chosen sets for 44kHz and to 22.9% of the chosen sets for 8kHz. This illustrates that using prior information for the feature choice is a good approach to follow. However, looking at the distribution of the chosen features by category (Figures 5.3 and 5.4) one can see that attention must be payed to other features, as audSpec and pcm. In further research, these may belong to another Informed Choice Set.

Though features were chosen for the quality of classification they provide within a certain set, these frequencies may indicate special discriminating power on these types of features, since at least the first one to be chosen is chosen only by its individual discriminating power.

The evaluation of the performance of the classifier on the whole initial set gives good insight on the difference provided by different kernels. SVM with linear kernel classified with recognition rate 0.0 Feature Sets I and I+II, while SVM with gaussian kernel classified the same with recognition rate 0.492 . For all the considered combinations of classifiers and feature sets the was higher for reduced sets than for their corresponding complete sets (comparison of the two rightmost columns in Tables 5.5 and 5.6). These results corroborate the initial statement in this work that smaller sets of features provide better discrimination.

Only one TEO-Based feature was chosen, among all sets. This is unexpected, since in the literature TEO is shown to have a good performance in discriminating stress in speech. Several aspects may have contributed for this. The first one is the fact that the wrapper algorithm of choice stops at a local minimum, not a global minimum, which means that search is done not for the best feature set, but for the best feature set found immediately before its rate cannot be increased by any feature else. Therefore, it is possible that these features are still helpful to discriminate stress, even if not chosen by the methods used here – Guyon and Elisseeff (2003) explain that apparently innocuous features can help to provide good discrimination if rightly combined. The

Table 5.6: Performance of the Selected Feature Sets, sampling at 8kHz. SF is the Sampling Frequency at which audio files were recorded. Train. RR is the Recognition Rate obtained in the feature selection. Gen. RR is the Recognition Rate in the test set. CS. RR is the Recognition Rate when training and testing on the full feature set.

Feature Set	**kernel**	**# Features**	**Train. RR**	**Gen. RR**	**CS RR**
Feature Set I	linear	8	0.745	0.670	0.0
Informed Choice Feature Set	linear	2	0.709	0.616	0.568
Feature Set I + II	linear	5	0.728	0.595	0.0
Feature Set I	gaussian	11	0.881	0.735	0.492
Informed Choice Feature Set	gaussian	2	0.709	0.616	0.492
Feature Set I + II	gaussian	7	0.849	0.714	0.492

second of these aspects is the way these features were implemented: though following the information in the literature as possible, as stated when describing their implementation (v. Subsection 4.4.6), this implementation could not happen exactly as described, which may also lead to some decrease in their discriminating power. Other aspect that may be considered is the range of situations where TEO-Based features are said to perform well. For instance, Jabloun et al. (1999) uses TEO to improve speech recognition in the presence of car engine noise, and in Lu et al. (2012) TEO-Based features provide the best stress discrimination only in outdoor activities. These factors represent quality improvement in noisy environments. Now, in VOCE experimental set, no recordings are *a priori* more noisy than others. Therefore, it is possible that this is not the context where TEO-Based features provide more powerful discrimination than other features. Furthermore, this combination of several phones into a single data point (utterance) has not been seen in literature, where phones are analysed one at a time. The removal of many voiced phones, due to the condition of being vowels, also may have lead to a considerable loss of data.

Recognition rates were higher for features extracted from 44kHz than from 8kHz. On average, recognition rates on generalisation are 0.698 for 44kHz and 0.658 for 8kHz. The average difference between recognition rates for the two sampling frequencies, and for the corresponding feature sets and classifiers is 0.040. This can be due to the missing information on parts of the spectrum that are considered at 44kHz and not at 8kHz. Zhou et al. (2001) claims that under stress there is a shift of the formant frequencies to other frequency bands. It is possible that those fine effects are better captured and noticed by considering a higher sampling frequency. Nevertheless, the results obtained for 8kHz help us to understand how it can be implemented in a real system, where sampling at 44kHz is not an option. These performance differences also allow us to know the discrimination power that can be lost for doing classification in a real system. Furthermore, considering that the size of test set is less than half of the training set it can be said that the loss in precision due to the difference in sampling frequency is very small. The loss in recognition rate obtained from 44kHz to 8kHz is very small, which is a good result, considering the referred possible losses and the goal of this research.

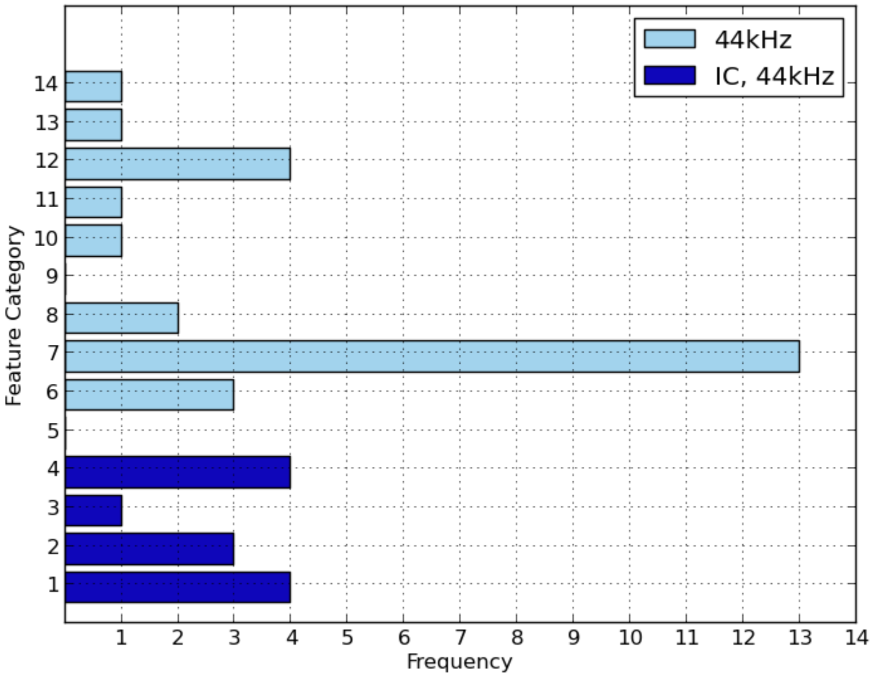

Figure 5.3: Frequency of selection on feature types, at 44kHz.1 - voicingFinalUnclipped_sma, 2 - jitterLocal_sma, 3 - jitterDDP_sma, 4 - shimmerLocal_sma, 5 - logHNR_sma, 6 - Critical_band_based_TEO_autocorrelation_envelope_ band16, 7 - F0final_sma, 8 - audSpe_Rfilt, 9 - audspecRasta_lengthL1norm, 10 - mfcc_sma, 11 - pcm_Mag, 12 - pcm_RMSenergy, 13 - pcm_zcr, 14 - voicingFinalUnclipped_sma .

Table 5.7: Recognition rates in Literature

Work	**Precision**	**# Features**
Patil and Hansen (2008)	0.6 to 0.8	37
Lu et al. (2012)	0.81 indoor and 0.76 outdoor	42
Fernandez and Picard (2003)	0.512 %	10

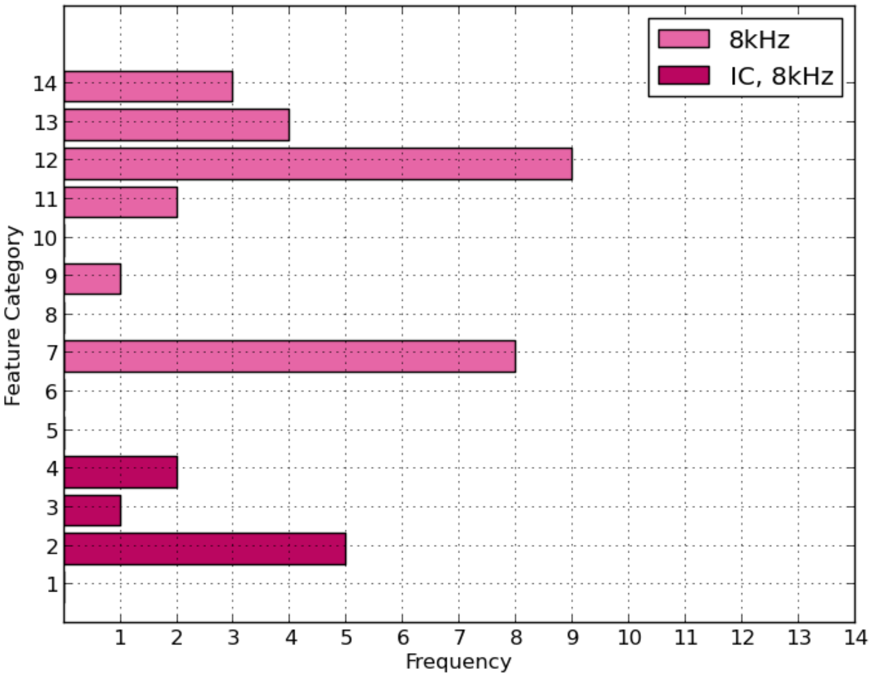

Figure 5.4: Frequency of selection on feature types, at 8kHz.1 - voicingFinalUnclipped_sma, 2 - jitterLocal_sma, 3 - jitterDDP_sma, 4 - shimmerLocal_sma, 5 - logHNR_sma, 6 - Critical_band_based_TEO_autocorrelation_envelope_ band16, 7 - F0final_sma, 8 - audSpe_Rfilt, 9 - audspecRasta_lengthL1norm, 10 - mfcc_sma, 11 - pcm_Mag, 12 - pcm_RMSenergy, 13 - pcm_zcr, 14 - voicingFinalUnclipped_sma .

Chapter 6

Conclusions and Future Work

6.1 Goal Achievement

In Section 1.3, it is defined as objective to understand which of the available feature sets provides the best performance for stressed/neutral discrimination. Results showed us that the sets providing best performance were Reduced Feature Set I + II for SVM with linear kernel, for audio recordings sampled at 44.1kHz, and Reduced Feature Set I for SVM with gaussian kernel, for audios sampled at 8kHz with recognition rates 0.741 and 0.735, respectively. These results are not particularly better than others : average recognition rates (in generalization) are 0.698 for 44kHz and 0.658 for 8kHz. These results are very good considering not only they are within results commonly seen in literature, but also because they are completely speaker-independent – utterances used for test were obtained from speakers whose utterances were not used for train.

In all the evaluated cases subsets of the main set allowed for more accurate discrimination than whole sets, which corroborated the assumption that dimensionality reduction leads to more accurate results.

Comparing the best results for both analysed sampling frequencies, the reduction in the recognition rate was of only of 0.006. This is very encouraging, since it means the loss of acuracy for implementation in real systems is almost neglectable.

As previously said, for Feature Set I and Feature Set I+II, chosen features belonging to the Informed Choice Feature Set correspond to 33.2% and 22.9% of the total amount of features for 44kHz and 8kHz, respectively. This indicates that these features are good stress discriminators, thus reinforcing the idea of using *a priori* knowledge for feature selection.

6.2 Future Work

During the development of this work some questions arose whose study could lead to a fuller accomplishment of the goal of this thesis.

One of these is the way classification is done. So far, recognition rates only considered the rate of correctly classified utterances. In the future, attention must be payed to which the correctly and wrongly classified utterances are: if stressed utterances tend to be correctly classified in the same way as neutral ones, or if recognition rates are the same for male and female speakers.

Concerning TEO, it may be that better results are obtained by considering some aspects. First of all, it would be interesting to understand the misbehaviour in TEO-based "cepstral coefficients", that led it to unsuitable results. Then, it would be good to fully comprehend thwarts in the process of phone selection and feature extraction. The knowledge of its main sources (duration and noise) was not enough to ensure reliability in the process. Therefore, they shall be understood in such a way that the process of feature extraction can be made more straightforward and reliable. In addition to these, voiced but not-vowel phones can also be considered for this extraction, since they also are voiced sounds. Meta-features, as voice breaks or speaking rate shall also be brought into the feature set.

Since it was seen that slightly different classifiers can behave very differently too, the idea of using ensemble learning is also very atractive. Classifiers may not be equally adequate to all types of features and, consequently a combination of simple classifiers may lead to an improvement of the results.

The problem with wrapper algorithm is to stop in a local minimum. One way to go around local minima is to use meta-heuristics like simulated annealing, that try to avoid them. Another way to go around this problem, though a simpler one, is to change the stopping criterion, by imposing the algorithm to stop after some deliberated number of features have been chosen. In practice, a feature would always be appended to the previous set, even though the rate would not be improved. This would be the feature for which that feature set would return the greatest recognition rate. Then, within that number of possibly chosen feature sets it would be chosen the feature set at the point of returning the best recognition rate.

Even though starting from different sets, the resulting feature sets, for being extracted from highly overlapping sets, leave us with the idea that shuffling and randomly sampling the feature sets before applying feature selection would be a good approach to the feature selection process. This is close to the concept of "bootstraps", where feature selection occurs in subsamples of the data, and the union of the selected features works as the final selected set (Guyon and Elisseeff (2003)), and may be a good path to follow for future research.

References

Alberto Abad. The l2f language recognition system for albayzin 2012 evaluation. In *In Proceedings of IberSPEECH*, 2012.

Ana Aguiar, Mariana Kaiseler, Hugo Meinedo, Pedro Almeida, Mariana Cunha, and Jorge Silva. Voce corpus: Ecologically collected speech annotated with physiological and psychological stress assessments. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland, May 2014. European Language Resources Association (ELRA). ISBN 978-2-9517408-8-4.

Ana C. Aguiar, Mariana Kaiseler, Hugo Meinedo, Traian E. Abrudan, and Pedro R. Almeida. Speech stress assessment using physiological and psychological measures. In Friedemann Mattern, Silvia Santini, John F. Canny, Marc Langheinrich, and Jun Rekimoto, editors, *UbiComp (Adjunct Publication)*, pages 921–930. ACM, 2013. ISBN 978-1-4503-2215-7.

Rainer Banse and Klaus R. Scherer. Acoustic profiles in vocal emotion expression. *Journal of Personality and Social Psychology*, 70(3):614–636, 1996.

Fernando Batista, Helena Moniz, Isabel Trancoso, and Nuno J. Mamede. Bilingual experiments on automatic recovery of capitalization and punctuation of automatic speech transcripts. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(2):474–485, 2012.

Anton Batliner, Stefan Steidl, Björn Schuller, Dino Seppi, Kornel Laskowski, Thurid Vogt, Laurence Devillers, Laurence Vidrascu, Noam Amir, Loic Kessous, and Vered Aharonson. Combining efforts for improving automatic classification of emotional user states. In *Proc. IS-LTC 2006, Ljubliana*, pages 240–245, 2006.

Sahar E. Bou-Ghazale and John H. L. Hansen. Generating stressed speech from neutral speech using a modified celp vocoder. *Speech Commun.*, 20(1-2):93–110, November 1996. ISSN 0167-6393. doi: 10.1016/S0167-6393(96)00047-7. URL http://dx.doi.org/10.1016/S0167-6393(96)00047-7.

Sahar E. Bou-Ghazale and John H. L. Hansen. Hmm-based stressed speech modeling with application to improved synthesis and recognition of isolated speech under stress. *IEEE Transactions on Speech and Audio Processing*, 6(3):201–216, 1998. URL http://dblp.uni-trier.de/db/journals/taslp/taslp6.html#Bou-GhazaleH98.

Sahar E. Bou-Ghazale and John H. L. Hansen. A comparative study of traditional and newly proposed features for recognition of speech under stress. *IEEE Trans. Speech Audio Process.*, 8(4):429–442, July 2000.

Douglas A. Cairns, John H. L. Hansen, and James F. Kaiser. Recent advances in hypernasal speech detection using the nonlinear teager energy operator. In *ICSLP'96*, pages –1–1, 1996.

Zetao Chen and Xin Li. Analysis of stress rating model based on hmm, 2013. URL www.ori-medsci.com/en/.

Renee Peje Clapham, Lisette van der Molen, R. J. J. H. van Son, Michiel W. M. van den Brekel, and Frans J. M. Hilgers. Nki-ccrt corpus - speech intelligibility before and after advanced head and neck cancer treated with concomitant chemoradiotherapy. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Mehmet Ugur Dogan, Bente Maegaard, Joseph Mariani, Jan Odijk, and Stelios Piperidis, editors, *LREC*, pages 3350–3355. European Language Resources Association (ELRA), 2012. ISBN 978-2-9517408-7-7. URL http://dblp.uni-trier.de/db/conf/lrec/lrec2012.html#ClaphamMSBH12.

Grazyna Demenko. Voice stress extraction. *Proceedings of the Speech Prosody 2008 Conference*, 2008.

Grazyna Demenko and Magdalena Jastrzebska. Analysis of voice stress in call centers conversations. *Proc. of Speech Prosody, 6th International Conference, Shanghai, China*, 2012.

Ellen Douglas-Cowie, Nick Campbell, Roddy Cowie, and Peter Roach. Emotional speech: Towards a new generation of databases, 2003.

I.S. Engberg and A.V. Hansen. Documentation of the danish emotional speech database (des), 1996.

Florian Eyben, Martin Wöllmer, and Björn Schuller. Opensmile: the munich versatile and fast open-source audio feature extractor. In Alberto Del Bimbo, Shih-Fu Chang, and Arnold W. M. Smeulders, editors, *ACM Multimedia*, pages 1459–1462. ACM, 2010. ISBN 978-1-60558-933-6. URL http://dblp.uni-trier.de/db/conf/mm/mm2010.html#EybenWS10.

Raul Fernandez and Rosalind W. Picard. Modeling drivers' speech under stress. *Speech Communication*, 40(1-2):145–159, 2003.

Jaime Ferreira and Hugo Meinedo. Voce project stress feature survey technical report 2. Technical report, L2F, Inesc-ID, Lisbor, Portugal, November 2013.

Isabelle Guyon and André Elisseeff. An introduction to variable and feature selection. *Journal of Machine Learning Research*, 3:1157–1182, 2003.

John H. L. Hansen. Analysis and compensation of speech under stress and noise for environmental robustness in speech recognition. *Speech Commun.*, 20:151–173, Nov. 1996.

John H. L. Hansen, Wooil Kim, Mandar Rahurkar, Evan Ruzanski, and James Meyerhoff. Robust emotional stressed speech detection using weighted frequency subbands. *EURASIP J. Adv. Sig. Proc.*, 2011, 2011. URL http://dblp.uni-trier.de/db/journals/ejasp/ejasp2011.html#HansenKRRM11.

John HL Hansen and Sanjay A. Patil. Speech under stress: Analysis, modeling and recognition, 2007.

John H.L. Hansen, Sahar E. Bou-Ghazale, Ruhi Sarikaya, and Bryan Pellom. Getting started with the susas: A speech under simulated and actual stress database. *Technical Report: RSPL-98-10*, 1998.

Firas Jabloun, A. Enis Çetin, and Engin Erzin. Teager energy based feature parameters for speech recognition in car noise. *IEEE Signal Processing Letters*, 1999.

J.F. Kaiser. On a simple algorithm to calculate the 'energy' of a signal. *Proc. ICASSP, Albuquerque, NM*, 1990a.

J.F. Kaiser. On teager's energy algorithm and its generalization to continuous signals. *Proceedings of IEEE DSP Workshop*, 1990b.

G. U. Kharat and S. V. Dudul. Human emotion recognition system using optimally designed svm with different facial feature extraction techniques. *W. Trans. on Comp.*, 7(6):650–659, June 2008. ISSN 1109-2750. URL http://dl.acm.org/citation.cfm?id=1458369.1458376.

Kenji Kira and Larry A. Rendell. A practical approach to feature selection. In *Proceedings of the Ninth International Workshop on Machine Learning*, ML92, pages 249–256, San Francisco, CA, USA, 1992. Morgan Kaufmann Publishers Inc. ISBN 1-5586-247-X. URL http://dl.acm.org/citation.cfm?id=141975.142034.

Ron Kohavi and George H. John. Wrappers for feature subset selection. *ARTIFICIAL INTELLIGENCE*, 97(1):273–324, 1997.

Eivind Kvedalen. *Signal processing using the Teager Energy Operator and other nonlinear operators*. PhD thesis, University of Oslo, Department of Informatics, 2003.

Y Li and Y Zhao. Recognizing emotions in speech using short-term and long-term features, 1998. Proceedings of ICSLP98.

YL. Lin and G. Wei. Speech emotion recognition based on hmm and svm, 2005.

Hong Lu, Denise Frauendorfer, Mashfiqui Rabbi, Marianne Schmid Mast, Gokul T. Chittaranjan, Andrew T. Campbell, Daniel Gatica-Perez, and Tanzeem Choudhury. Stresssense: Detecting stress in unconstrained acoustic environments using smartphones. In *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*, UbiComp '12, pages 351–360, New York, NY, USA, 2012. ACM. ISBN 978-1-4503-1224-0. doi: 10.1145/2370216.2370270. URL http://doi.acm.org/10.1145/2370216.2370270.

Petros Maragos, James F Kaiser, and Thomas F Quatieri. Energy separation in signal modulations with application to speech analysis, 1993.

Nobuaki Minematsu, Satoshi Kobashikawa, Keikichi Hirose, and Donna Erickson. Acoustic modeling of sentence stress using differential features between syllables for english rhythm learning system development. In *7th International Conference on Spoken Language Processing, ICSLP2002 - INTERSPEECH 2002, Denver, Colorado, USA, September 16-20, 2002*, 2002. URL http://www.isca-speech.org/archive/icslp_2002/i02_0745.html.

N. Sundaram N., B.Y. Smolenski, and R. Yantorno. Instantaneous nonlinear teager energy operator for robust voiced–unvoiced speech classification, 2003. URL http://www.temple.edu/speech_lab/sundaram.PDF.

Sanjay A. Patil and John H. L. Hansen. Detection of speech under physical stress: Model development, sensor selection, and feature fusion. In *Proc. Interspeech*, 2008.

Robert Ruiz, Emmanuelle Absil, Bernard Harmegnies, Claude Legros, and Dolors Poch. Time- and spectrum-related variabilities in stressed speech under laboratory and real conditions. *Speech Communication*, 20(1-2):111–129, 1996. URL http://dblp.uni-trier.de/db/journals/speech/speech20.html#RuizAHLP96.

Evan Ruzanski, John H. L. Hansen, James Meyerhoff, George Saviolakis, and Michael Koenig. Effects of phoneme characteristics on teo feature-based automatic stress detection in speech. In *ICASSP (1)*, pages 357–360, 2005.

Ruhi Sarikaya and John N. Gowdy. Subband based classification of speech under stress. In *ICASSP*, pages 569–572, 1998.

Klaus R. Scherer. Vocal communication of emotion: A review of research paradigms. *Speech Commun.*, 40(1-2):227–256, April 2003. ISSN 0167-6393. doi: 10.1016/S0167-6393(02)00084-5. URL http://dx.doi.org/10.1016/S0167-6393(02)00084-5.

Klaus R. Scherer, Didier Grandjean, Tom Johnstone, Gudrun Klasmeyer, and Thomas Bänziger. Acoustic correlates of task load and stress. In John H. L. Hansen and Bryan L. Pellom, editors, *INTERSPEECH*. ISCA, 2002.

Stefan Scherer, Hansjörg Hofmann, Malte Lampmann, Martin Pfeil, Steffen Rhinow, Friedhelm Schwenker, and Günther Palm. Emotion recognition from speech: Stress experiment. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, and Daniel Tapias, editors, *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, may 2008. European Language Resources Association (ELRA). ISBN 2-9517408-4-0. http://www.lrec-conf.org/proceedings/lrec2008/.

Björn Schuller, Anton Batliner, Stefan Steidl, and Dino Seppi. Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge. *Speech Commun.*, 53(9-10):1062–1087, November 2011. ISSN 0167-6393. doi: 10.1016/j.specom.2011.01.011. URL http://dx.doi.org/10.1016/j.specom.2011.01.011.

Björn Schuller, Anton Batliner, Dino Seppi, Stefan Steidl, Thurid Vogt, Johannes Wagner, Laurence Devillers, Laurence Vidrascu, Noam Amir, Loïc Kessous, and Vered Aharonson. The relevance of feature type for the automatic classification of emotional user states: low level descriptors and functionals. In *INTERSPEECH*, pages 2253–2256. ISCA, 2007. URL http://dblp.uni-trier.de/db/conf/interspeech/interspeech2007.html#SchullerBSSVWDVAKA07.

Björn Schuller, Matthias Wimmer, Lorenz Moesenlechner, Christian Kern, Dejan Arsic, and Gerhard Rigoll. Brute-forcing hierarchical functionals for paralinguistics: A waste of feature space? In *ICASSP*, pages 4501–4504. IEEE, 2008. ISBN 1-4244-1484-9. URL http://dblp.uni-trier.de/db/conf/icassp/icassp2008.html#SchullerWMKAR08.

Björn Schuller, Stefan Steidl, Anton Batliner, Elmar Nöth, Alessandro Vinciarelli, Felix Burkhardt, Rob van Son, Felix Weninger, Florian Eyben, Tobias Bocklet, Gelareh Mohammadi, and Benjamin Weiss. The interspeech 2012 speaker trait challenge. In *INTERSPEECH*. ISCA, 2012.

Nandita Sharma and Tom Gedeon. Objective measures, sensors and computational techniques for stress recognition and classification: A survey. *Comput. Methods Prog. Biomed.*, 108(3):1287–1301, December 2012. ISSN 0169-2607. doi: 10.1016/j.cmpb.2012.07.003. URL http://dx.doi.org/10.1016/j.cmpb.2012.07.003.

Julius O. Smith and Jonathan S. Abel. Bark and erb bilinear transforms. *IEEE Transactions on Speech and Audio Processing*, 7:697–708, 1999.

Xuejing Sun. A pitch determination algorithm based on subharmonic-to-harmonic ratio. In *the 6th International Conference of Spoken Language Processing*, pages 676–679, 2000.

H. M. Teager. Some observations on oral air flow during phonation. *Acoustics, Speech, Signal Processing, IEEE Transactions on*, 1980.

H. M. Teager and S.M. Teager. A phenomenological model for vowel production in the vocal tract. *Speech Science: Recent Advances*, pages 73–109, 1985.

H. M. Teager and S.M. Teager. Evidence for nonlinear sound production mechanisms in the vocal tract. *France: Kluwer Acad. Publ.*, pages 241–261, 1990.

Dimitrios Ververidis and Constantine Kotropoulos. A review of emotional speech databases, 2003.

Dimitrios Ververidis and Constantine Kotropoulos. Emotional speech classification using gaussian mixture models and the sequential floating forward selection algorithm. In *Multimedia and Expo, 2005. ICME 2005. IEEE International Conference on*, pages 1500–1503. IEEE, 2005.

Dimitrios Ververidis and Constantine Kotropoulos. Emotional speech recognition: Resources, features, and methods. *Speech Communication*, 48(9):1162–1181, 2006. URL http://dblp.uni-trier.de/db/journals/speech/speech48.html#VerveridisK06.

Thurid Vogt, Elisabeth André, and Johannes Wagner. Automatic recognition of emotions from speech: a review of the literature and recommendations for practical realisation. In *In LNCS 4868*, pages 75–91, 2008.

Karl E. Williams and Stevens N. Kenneth. Emotions and speech: Some acoustical correlates. *Proceedings of the Speech Prosody 2008 Conference*, 1972.

N. Yang, R. Muraleedharan, J. Kohl, I. Demirkol, W. Heinzelman, and M. Sturge-Apple. Speech-based emotion classification using multiclass svm with hybrid kernel and thresholding fusion. In *SLT*, pages 455–460. IEEE, 2012. ISBN 978-1-4673-5125-6. URL http://dblp.uni-trier.de/db/conf/slt/slt2012.html#YangMKDHS12.

Guojun Zhou, J.H.L. Hansen, and J.F. Kaiser. Nonlinear feature based classification of speech under stress. *Speech and Audio Processing, IEEE Transactions on*, 9, 2001.