That voice sounds familiar

Factors in speaker recognition

Erik J. Eriksson
ABSTRACT

Humans have the ability to recognize other humans by voice alone. This is important both socially and for the robustness of speech perception. This Thesis contains a set of eight studies that investigates how different factors impact on speaker recognition and how these factors can help explain how listeners perceive and evaluate speaker identity. The first study is a review paper overviewing emotion decoding and encoding research. The second study compares the relative importance of the emotional tone in the voice and the emotional content of the message. A mismatch between these was shown to impact upon decoding speed. The third study investigates the factor dialect in speaker recognition and shows, using a bidialectal speaker as the target voice to control all other variables, that the dominance of dialect cannot be overcome. The fourth paper investigates if imitated stage dialects are as perceptually dominant as natural dialects. It was found that a professional actor could disguise his voice successfully by imitating a dialect, yet that a listener’s proficiency in a language or accent can reduce susceptibility to a dialect imitation. Papers five to seven focus on automatic techniques for speaker separation. Paper five shows that a method developed for Australian English diphthongs produced comparable results with a Swedish glide + vowel transition. The sixth and seventh papers investigate a speaker separation technique developed for American English. It was found that the technique could be used to separate Swedish speakers and that it is robust against professional imitations. Paper eight investigates how age and hearing impact upon earwitness reliability. This study shows that a senior citizen with corrected hearing can be as reliable an earwitness as a younger adult with no hearing problem, but suggests that a witness’ general cognitive skill deterioration needs to be considered when assessing a senior citizen’s earwitness evidence. On the basis of the studies a model of speaker recognition is presented, based on the face recognition model by V. Bruce and Young (1986; British Journal of Psychology, 77, pp. 305 – 327) and the voice recognition model by Belin, Festag and Bédard (2004; TRENDS in Cognitive Science, 8, pp. 129 – 134). The merged and modified model handles both familiar and unfamiliar voices. The findings presented in this Thesis, in particular the findings of the individual papers in Part II, have implications for criminal cases in which speaker recognition forms a part. The findings feed directly into the growing body of forensic phonetic and forensic linguistic research.
ACKNOWLEDGEMENT

The research reported in this thesis were undertaken within the project Imitated Voices: a research project with applications to security and the law (Dnr: K2002-1121:1–4) which was funded by The Bank of Sweden Tercentenary Fund. The fund and the project leader Kirk Sullivan are hereby acknowledged.

I owe my supervisor Kirk Sullivan many thanks. First, for making it possible for me to undergoe my PhD education to completions. Second, for taking the time to co-author and proof-reading the work included in this thesis. Third, for contributing to my personal development. Fourth, for, during long work hours, taking part in really quite unique and obviously really interesting discussions, really.

I also like to thank Prof. Robert Rodman at NCSU, NC, USA for hosting my research scholarship and providing a friendly and professionally excellent working environment. I also like to thank all the people at the Department of Computer Science, NCSU for providing the work place, materials, personal help and friendship during my stay in the USA.

I am grateful to all participants in the project in which this work was undertaken. Thank you Maria Sjöström and Tomas Landgren for analysing and preparing the data material and recording the participants. Thank you Elisabeth Zetterholm for providing recordings, ideas and comments. I also thank all the participants that were recorded for the database developed.

I am also grateful to the people at the Department of Philosophy and Linguistics at Umeå University. Thank you Görel Sandström for reading the draft of my thesis, and for discussing matters marginally related to work during lunch and coffee breaks. Thank you also Fredrik Karlsson for the discussions during, sometimes too long, coffee breaks. I also thank Felix Schaeffler for good and fun discussions and thank you Leila Kantola for being the verbal APA manual. I extend my deepest gratitude to Jan van Doorn for providing housing during my stay in Australia and organizing the beer committee; it must prevail.

Finally, I would like to thank my family who put up with the stress and constant nagging when times were hard and the inexplainable euphoria when times were good. I especially thank Karolina for taking care of me and Hilding. I also thank my parents, Gunilla and Johannes, for offering their support unconditionally.
CONTENTS

List of papers 3

1 Introduction 4

2 Speaker recognition background 4
  2.1 Behavioural evidence .................................................. 5
  2.2 Neurological evidence .................................................. 7

3 Methods 8
  3.1 Speaker similarity judgements ................................. 9
  3.2 Voice line-ups ......................................................... 9

4 Acoustic and perceptual factors in speaker recognition 10
  4.1 Evaluative factors .................................................. 10
    4.1.1 Gender ................................................................. 10
    4.1.2 Regional dialect .................................................. 11
    4.1.3 Foreign accents .................................................. 12
    4.1.4 Age ................................................................. 13
    4.1.5 Distinctiveness .................................................. 14
    4.1.6 Disguise .............................................................. 14
    4.1.7 Emotions ............................................................ 14
  4.2 Measureable factors .................................................. 14
    4.2.1 Formant transitions ........................................... 15
    4.2.2 Fundamental Frequency ....................................... 17
    4.2.3 LTAS ................................................................. 17
  4.3 External factors ...................................................... 17
    4.3.1 Retention interval ............................................... 17
    4.3.2 Sample duration and quality .................................. 18
    4.3.3 Speaker familiarity .............................................. 18
  4.4 Factor summary ...................................................... 19

5 Materials and Papers 20
  5.1 UDID – Umeå disguise and imitation database ................. 20
  5.2 Summary of Papers .................................................. 22
    5.2.1 Paper 1 – Emotions in Speech: Judicial Implications .... 22
    5.2.2 Paper 2 – Acoustic Impact on Decoding of Semantic Emotion 22
    5.2.3 Paper 3 – On the perceptual dominance of dialect .......... 23
    5.2.4 Paper 4 – Dialect imitations in speaker identification .... 23
    5.2.5 Paper 5 – An investigation of the effectiveness of a Swedish glide + vowel segment for speaker discrimination 23
    5.2.6 Paper 6 – Cross-language speaker identification using spectral moments .......................... 24
5.2.7 Paper 7 – Robustness of Spectral Moments: a Study using Voice Imitations 24
5.2.8 Paper 8 – Effects of age and age-related hearing loss on speaker recognition or can senior citizens be reliable earwitnesses 25

6 Memory models of speaker recognition 25
6.1 Pattern recognition model 26
6.2 Neurological model of speaker recognition 27
6.3 Prototype model of speaker identification 29
6.4 Discussion 30

7 Conclusion 32

8 Suggested areas for future research 33

References 35
LIST OF PAPERS


1. INTRODUCTION

All listeners have experienced recognition of a person by a short verbal presentation alone. The person that is identified is often highly familiar and the context surrounding the identification is also often specific. A voice may be heard from a television and the identity of the speaker is recognised, almost automatically (Hollien, 2002). Here, the context is the television and probably a specific programme. Similar effects can be found when telephoning a relative. The effect of recognition is more often noticed when it fails, e.g. when a call to relative is misdialled or when someone unexpected answers the phone. Conversely, we may find ourselves having spoken, sometimes at length, to someone that is misrecognized as someone else.

The process of speaker identification is complex and integrated into other processes. For instance, Pisoni (1997) suggested that speaker identity and item memory (i.e. memory for words and sentences) are integrated and dependent on one another. However, single utterances containing little or no linguistic message can still lead to speaker identification (Hollien, 2002). What in the voice makes it memorable is thus an outstanding and important question. As Carterette and Barnebey (1975) argued:

If a voice which will later be heard again is stripped of its semantical, grammatical and contextual constraints so as to lose its specialness of speech except as a carrier, are its abstracted properties laid down in a speech code or a memory code? The answer is important in the biology of survival, and also in our own human society which is held together by voice communication. The day is near when men and machines will talk fluently to each other. And even if it were not, the answer is interesting because whatever the evidence for or against, it is widely held that a voice can be recognized as familiar from a brief fragment of speech. (p. 246)

The work presented in this Thesis deals with some variables involved in speaker recognition. The thesis is set up in two parts. Part I gives a background to the theories of speaker recognition as well as a summary of the papers in Part II and concludes with a discussion about the relevance of these studies.

2. SPEAKER RECOGNITION BACKGROUND

The premise in speaker identification is that there exists a set of variables for a speaker, such that these variables, taken together, abstracting away from the linguistic content of the message, define this speaker uniquely.
The framework applies to naïve listeners as well as experts or trained evaluators. This framework is in a sense the opposite of traditional phonetics, which investigates speech phenomena pertaining to a language or variant, abstracting away from speaker variation within the community. As van Dommelen (1990) put it (p. 259):

... if we focus on the information for signalling accents coded in an F0 contour, our quest concerns features which belong to the speech code and which are common to all speakers of a speech community. Inter-speaker variability is considered an inevitable artefact that should be eliminated as far as possible.

In traditional phonetics, with its focus on the invariant code as the goal of the analysis, the speaker dependent variation becomes a problem. The process through which listeners are able to recover invariance has been termed speaker normalization and functions as a way of reducing the variation in the source so linguistic content can be evaluated (see Pisoni, 1997 and Goldinger, 1996 for an overview of the history of normalization research). Normalization theory, including its need to separate the signal into linguistic content and speaker identification content, has been questioned (see, among others Pisoni, 1997). Pisoni argued that the results that purport to support the normalization process came from small sample sets with few speakers. More recent speech databases include larger sample set and, inherently, more variation. Further, Pisoni argued that this variation is essential to the perception system, otherwise correct decoding of linguistic information would be impossible in less-than-perfect circumstances. The speaker dependent information carried by a voice that not only defines a speaker’s identity, or contributes to the recognition of that speaker, but is also an important factor in speech perception.

2.1 Behavioural evidence

Many studies have examined the impact of speaker variation on speech decoding and representation (e.g. Mullennix & Pisoni, 1990; Remez, Fellowes, & Rubin, 1997; Sheffert, Pisoni, Fellowes, & Remez, 2002). These have shown that speaker information in its minute phonetic detail is not only encoded simultaneously as linguistic information but is also a vital part of the memory representation of speech segments. Mullennix and Pisoni investigated identification of word-initial consonants by manipulation of the speaker identity. In their speeded classification task listeners were unable to attend to one of the two dimensions selectively when the two dimensions were manipulated simultaneously; "Information about word-initial consonants and information about the talker’s
voice appear to be processed together in a mutually dependent manner” (p. 385). When manipulating one dimension at a time, similar effects were found. That is, attention to the target dimension (either consonant classification or speaker classification) was detrimentally affected by variation in the other dimension. However, Mullennix and Pisoni also found that the two dimensions did not function equally with respect to impact on processing load. Voice variation initially had, as the number of voices increased, an increased effect on word classification but this effect levelled off; the effect for word variation on voice classification was linearly related to the number of words presented, with a steady increase in processing cost.

Several studies have shown the impact of gross speaker characteristics, specifically gender, on sentence recognition (Geiselman & Bellezza, 1976, 1977). Geiselman and Bellezza (1976) argued for a voice connotation model, where the voice characteristics are encoded as “an integral part of the code” (Geiselman & Bellezza, 1977, p. 659). Geiselman and Bellezza (1977) argued that gender was a part of the stimuli which was processed simultaneously with the linguistic content of the stimuli.

Palmeri, Goldinger, and Pisoni (1993) tested listeners’ recognition of previously presented words in lists to investigate the impact of speaker variation on word recall. The results indicated an impact of speaker identity on word recognition speed. That is, if the word to be recalled was presented by the same speaker as during initial encoding phase the word was recognized both faster and more accurately than if the word was presented by different speakers in the encoding and recall phases. Further, when presented with lists read by two different speakers, listeners were accurate in judging whether the recalled item was spoken by the same speaker as during encoding or a different one. For word lists read by more than two different speakers listeners showed a tendency to group speakers in terms of gender. That is, even though the speakers were different but of the same gender, listeners judged them to be the same between the encoding and recall phase of the test. Palmeri et al. concluded that item memory carries more than just the linguistic content and that during judgement of speaker similarity (though not the primary task) listeners rely on speaker gender as the primary dimension of similarity.

Goldinger (1996) reinforced the finding by Geiselman and Bellezza (1976) of gender grouping by explicitly investigating the perceptual similarity between speakers before testing for voice impact on memory recall. He found that the effect of voice impact is increased by the perceptual distance between the voices. The perceptual distance, in turn, was primarily explained by gender, but even within gender, dissimilar voices exhibited the same impact upon memory recall.

The impact of voice identity on item memory was further investi-
gated by Goh (2005). He found that item memory was affected by voice familiarity and that this impact mainly affected the listeners’ response bias. That is, although listeners’ performance in the matching of stimuli declined if the stimulus was spoken by a different voice from that as presented during training, the listeners increased their false alarm rates when the stimulus was spoken by a previously presented voice. This meant that speaker identity can affect listeners’ ability to judge material as previously heard in situations when the material is new, though the speaker is not.

Speakers that are familiar (or previously presented), was shown to increase listener comprehensibility of spoken words (Nygaard & Pisoni, 1998); an effect not found when prestened with unfamiliar voices.

2.2 Neurological evidence

The effect of combined processing of both the linguistic and speaker information has been further confirmed in neurological studies. For instance, Kaganovich, Francis, and Melara (2006) used two tasks and two variations of the stimuli in each task to show similar degrading effects on performance in either task during cross-stimuli trials. They had listeners classify a sound as being either one of two vowels, ignoring speaker variation, or as produced by either one of two speakers, ignoring vowel variations. When the speaker was varied in the vowel classification task and when the vowel varied in the speaker classification task a performance loss was found based on the ignored dimension (filtering interference). The performance loss was equal between the two conditions. This loss in performance was accompanied by a sustained negativity after stimulus onset as measured by ERP, as early as 100ms after stimulus onset. This short reaction was also found by Knösche, Lattner, Maess, Schauer, and Friederici (2002) who argued that it showed that information types are processed in parallel and pre-attentively.

In section 2.1 it was suggested that speaker identity features are integrated with memory representations for linguistic content. Kaganovich et al. (2006) and Knösche et al. (2002) found evidence of parallel processing in the early auditory system of these two types. Other studies have found evidence supporting different neural paths for the analyses prior to encoding. For instance, Senkfor and van Petten (1998) and Wong, Nusbaum, and Small (2004) showed separate dissociated neural substrates for the process of linguistic content and speaker identity information.

Studies into voice activation of cortical regions have provided more information on dissociation of the processes involved in content integration. First, a number of areas that are associated with voice, and voice alone, have been found (Belin, Zatorre, & Ahad, 2002; Belin, Zatorre,
Lafaille, Ahad, & Pike, 2000; Kriegstein & Giraud, 2003; Kriegstein, Eger, Kleinschmidt, & Giraud, 2003; Stevens, 2004). Second, within these areas more selective regions have been discovered: differences in pitch and spectral processing (Zatorre, Evans, Meyer, & Gjedde, 1992), voice identity processing (Kriegstein & Giraud, 2003), naturalness of a voice (Lattner, Meyer, & Friederici, 2004), and voice familiarity (Beauchemin et al., 2006; D. Van Lancker & Kreiman, 1987; D. R. Van Lancker, Cummings, Kreiman, & Dobkin, 1988; D. R. Van Lancker, Kreiman, & Cummings, 1989; Shah et al., 2001).

The difference in processing of familiar voice versus non-familiar voices is prominent (D. Van Lancker, Kreiman, & Emmorey, 1985). D. Van Lancker and Kreiman (1987) further separated the functions of voice discrimination and voice recognition. The same regions involved in discrimination were later found to be associated with the processing of unfamiliar voices and the regions correlated with voice recognition primarily processed familiar voices (D. R. Van Lancker et al., 1989). The two processes relating to familiar and non-familiar voices (i.e. voice discrimination and voice recognition) were also found to be doubly dissociated (D. R. Van Lancker et al., 1988). D. R. Van Lancker et al. reported brain lesioned patients that could discriminate between voices, but not recognize highly familiar voices. They also found brain lesioned patients (differently lesioned areas than for the previous group) that could recognize familiar voices but could not separate two unfamiliar ones.

Thus it can be concluded that the processing of voice information is an integral part of speech perception and that although the processes pertaining to linguistic content and speaker recognizing are regionally separated they overlap and influence each other. Or, as Sheffert et al. (2002) put it (p. 1464): “… there is no single set of features or perceptual processes that can be used to identify both words and talkers.” However, only gender (Geiselman & Bellezza, 1977; Goldinger, 1996; Palmeri et al., 1993) or possibly relative F0 movement (Goldinger, 1996) have been demonstrated to be dimensions that explain the effects on speech perception. Therefore, as Palmeri et al. concluded that other dimensions of voice recognition and voice encoding need to be investigated.

3. METHODS

To test how speaker recognition and identification function and what impact these processes. Two main methods can be used. One, speaker similarity judgements can be used. Two, listeners can be asked to remember and later recall a specific voice in a voice line-up situation.
3.1 Speaker similarity judgements

Much of the data presented in the background sections of this Thesis are generated by methods that build on speaker similarity judgements. Listeners are, in these types of studies, asked to judge, either explicitly or implicitly, the similarity between speakers. Voices are presented to the listeners in pairs and listeners rate these on a, for instance, five-point (Remez, Wissig, Ferro, Liberman, & Landau, 2004), seven-point (Murry & Singh, 1980), or nine-point (Gelfer, 1993) scale. The magnitude of the impact of scale resolution differences on the overall results is currently not known (Kent, 1996).

Results from speaker similarity judgements must, however, be interpreted with care. The results of D. Van Lancker and Kreiman (1987), D. R. Van Lancker et al. (1988), and D. R. Van Lancker et al. (1989) show that speaker discrimination and speaker recognition are two distinct processes. This means that the effects of a specific feature found when judging speaker similarity may not be easily generalized to recognizing speakers. Gelfer (1993) argued that the method of correlating listeners’ similarity judgements with a set of acoustic or perceptual features is of limited use, given that the features selected are limited by the researcher’s preconceptions of what might be important and by the availability of reliable measures.

3.2 Voice line-ups

An alternative to measuring similarities is to have listeners learn and recognize speakers by their voice, and manipulate certain features in the speaker’s voice to investigate the impact of those features on listeners’ recognition of the speaker. Direct speaker identification research commonly uses voice line-ups as a method of collecting data about listeners’ accuracy in detecting speaker identity. The voice line-up is parallel in design to the visual line-up. However, it has been argued that the two types do not provide the same degree of accuracy (Yarmey, Yarmey, & Yarmey, 1994) or that there is even a theoretical argument that the two should function similarly (Hollien, 1996). The critique mainly targets the forensic application of the technique; as a tool to find criminals or suspects. It is used in research, where greater control of retention times and stimuli presentation is available, as a means to see how well a voice is recognized in a set of other voices. First, the target voice is presented to the listeners, often referred to as the familiarization phase or training phase, followed by a retention interval which may vary in length. The listeners are then asked to identify the voice they heard from a set of voices, the target may or may not be present in the line-up (closed or open sets of speakers). The data collection may be done in different ways: either the listeners are to respond with a number (e.g. Yarmey
et al., 1994), or answer yes or no to each presented voice (e.g. E. Eriksson, Kügler, Sullivan, van Doorn, & Zetterholm, 2003; Zetterholm et al., 2003).

4. ACOUSTIC AND PERCEPTUAL FACTORS IN SPEAKER RECOGNITION

In order to recognize a speaker a set of features delimiting the speaker’s identity must be available to the listener. Abercrombie (1967) argued for a set of indices that signalled information about the speaker, including regional and social group, age, and emotional state. These features should, therefore, be present in the acoustic signal and prominent to the listener. Further, the set of features should contain idiosyncratic information, which is information that is specific for a speaker.

Hollien (2002) presented a list of features that he claimed are used perceptually by listeners to identify a speaker. The list includes heard pitch, articulation, general voice quality, prosody, vocal intensity, and speech characteristics (segmental). This section presents factors that are related to speaker recognition and speaker identity.

4.1 Evaluative factors

The evaluative factors are factors that need interpretation by the listener. They are descriptive and are usually not linked to a specific set of measurable acoustic features.

4.1.1 Gender

Gender is a highly salient feature in the classification of voices (Cloppe & Pisoni, 2004a; Fellowes, Remez, & Rubin, 1997; Lass, Hughes, Bowyer, Waters, & Bourne, 1976; Murry & Singh, 1980). Lass et al. used recordings of speakers either speaking in their natural voice or whispering to investigate the impact of fundamental frequency on gender identification. They further included a low-pass filtered recording of the voiced samples. The results showed that listeners achieved best classification when the voiced recordings were played, slightly worse when the low-pass filtered stimuli were played and worst when the whispered stimuli were presented. However, Fellowes et al. showed that, by using sinewave replicas, listeners were able to detect speaker gender even though the fundamental frequency and vocal quality aspects were removed. Fellowes et al. additionally transposed each sinewave replica so that the gender information should be removed. Surprisingly, they found that listeners still were able to recognize individual speakers. That is, even though information about speaker gender should have been removed listeners could still identify speakers by their sinewave
replica, suggesting that listeners are able to use a multitude of features in their analysis of voice origin.

Murry and Singh (1980) investigated whether there were differences between similarity judgements for male and female voices presented by either a single sustained vowel or a whole sentence. They had listeners rate similarity between voices of both male and female speakers, but male and female voices were treated differently and were never matched to each other. They found that speaker gender influenced the set of parameters listeners used to evaluate speaker similarity. For male speakers the vowels and the sentences yielded similarity judgements that were correlated primarily with the measured fundamental frequency and perceived pitch and secondarily with “cues derived from vocal-tract resonance” (p. 296). For female voices, listeners used the fundamental frequency as the primary dimension when judging similarity between sustained vowels. However, when judging similarity between whole sentences Murry and Singh’s listeners primarily used the voice quality. Murry and Singh concluded that although gender is important in speaker discrimination and fundamental frequency is prominent, listeners may use different sets of features to distinguish between male speakers than to separate female speakers. Further, they argued that it may be that listeners primarily use voice quality to separate female speakers.

The different cues for different genders found by Murry and Singh (1980) were discussed by Gelfer (1993) who found that female speakers were judged as similar primarily by perceived pitch. She also found that, based on 17 different measures, voice quality had no great impact on similarity judgements of female speakers. She concluded that listeners do not use different sets of features to judge speaker similarity depending on gender.

4.1.2 Regional dialect
Regional dialect has been proposed as a signal of group membership (Abercrombie, 1967) and listeners ability to judge speakers’ regional origin based on voice alone has been investigated (Clopper & Pisoni, 2004b; Preston, 1993; Williams, Garrett, & Coupland, 1999). These results show that listeners are only able to classify speakers to a particular region with low regional resolution (Clopper & Pisoni, 2004b; Williams et al., 1999). Further, Preston (1993) showed that listeners’ background and knowledge of particular regional areas in the United States of America impacted upon their categorization of dialect regions. Remez et al. (2004) confirmed these findings by comparing similarity judgements of speakers from the same region and speakers from different regions evaluated by listeners with knowledge of one of the dialects but not the other. The results showed that listeners with knowledge of the regional dialect have a better resolution of speaker similarity than listeners that were
inexperienced with the dialect. This effect was also found for Dutch listeners and Dutch speakers in a voice line-up test (Kerstholt, Jansen, van Amelsvoort, & Broeders, 2006). Kerstholt et al. used two voices, one with a distinct Dutch dialect accent and one with a more standard Dutch dialect, and had listeners with and without experience of the distinct dialect accent respond to a speaker identification task. They found that the listeners were less able to identify the speaker of the distinct dialect than the speaker of the standard dialect. They concluded that exposure to the dialect impacted upon the listeners ability to detect speaker identity in the signal.

Finally, the distance between the dialect with which the listener is familiar and the dialect that the listener is to classify (Preston, 1993) or judge as similar (Clopper & Pisoni, 2004a) is related to the listeners’ resolution of the dialect presented. That is, the level of detail of dialect differences diminish with distance so that listeners tend to group speakers from large areas together in one group if the speakers’ dialect originate some distance away from the listener’s own dialect.

4.1.3 Foreign accents
Little research has been made on foreign accent in speaker identification. However, language awareness of the listener is one factor that has been related to the ability to separate speakers of another language (Schiller & Köster, 1996; Schiller, Köster, & Duckworth, 1997). Schiller and Köster (1996) investigated the impact of language awareness by letting three groups of listeners with different levels of experience in German take part in a speaker identification task. The groups were speakers of American English with no prior knowledge of German, a native English speaker with some experience in German, and a native speaker of German as control. The results show that an increased knowledge of the language increases the ability to identify speakers. They also found that the degree of knowledge of a language does not impact the ability to recognize speakers of that language; Schiller and Köster’s native German group and the group with some experience in German performed similarly. However, how knowledgeable a listener must be is not completely known. Sullivan and Schlichting (2000) found that British university BA-level students (after four years of study) were unable to attain the same level of performance as native speakers in a speaker identification test.

Doty (1998), reinforced the findings by Schiller and Köster (1996) and Schiller et al. (1997) by including several different nationalities and ethnicities in his analysis. He had speakers from the United States of America (USA), England, Belize and France, both male and females. For these nationalities the ethnicity varied: USA: African-American, Caucasian and Hispanic; England: Caucasian and Arabic-English; Belize: Creole,
Garifune, Latin, Mayan, Spanish and Mestizo; and France: Caucasians only. The listeners, on the other hand were controlled for age and living area, but not ethnicity. They were: USA: Hispanic, African-American and Caucasian; England: African-English, Middle Eastern-English and Caucasian. Each listener was exposed to a short excerpt from the target voice and then submitted to a voice line-up of ten voices each played consecutively. Each subject was exposed to two line-ups, one containing a male target and one a female target voice. Results showed that the listeners clearly identified speakers from their own country better than speakers from other countries. This was also true when the language was the same (i.e. English) but spoken with different accents (American or British); listeners were better at identifying speakers with their own country’s accent. For ethnicity among the listeners the only significant differences were that non-Caucasians were better at recognizing Belizean speakers than Caucasians.

Köster and Schiller (1997) used speakers of Chinese and Spanish with no or some knowledge in German to recognize German speakers. They found a difference, as detailed above, between native German speakers, speakers with some knowledge of German and speakers without knowledge of German. However, the typology (i.e. whether it was a tonal language or not) of the language did not affect the accuracy of the recognition.

4.1.4 Age

Abercrombie (1967) argued that age is something that affects the voice and therefore also can be detected and classified by listeners. In perceptual classification investigations it has been found that listeners are only able to assign speakers into broad age groups (e.g. Cerrato, Falcone, & Paoloni, 2000) and it depends on how the test is designed whether prediction of speaker age is successful or not (see Schötz, 2006, for a discussion). Further, it was argued by Braun (1996) that it is better to use age groups and classify speakers to that, or even only use descriptives such as ‘very young’ or ‘very old’.

In an experiment E. Eriksson, Green, Sjöström, Sullivan, and Zetterholm (2004) found, similarly to Braun (1996), that listeners over-estimate the chronological age of speakers, they rank them correctly. Thus, even if listeners are bad at specifically judging a speaker’s age based on voice alone, they are good at relationship judgements between speakers’ age. In comparison, Walden, Montgomery, Gibely, Prosek, and Schwartz (1978) used speaker similarity judgements between male voices and discovered that chronological age was highly correlated with the second psychological dimension explaining the most variance in the listeners’ similarity judgements.
4.1.5 Distinctiveness
The speaker’s specificity in the voice, or how the voice differs from other voices has also been proposed to be a function in speaker recognition. Yarmey (1991) argued that some speakers may be more distinct in their voice qualities so that they are more dissimilar to other voices whereas other speakers may be similar within a set. Yarmey (1991) defined the distinctiveness between speakers based on a set of features which included rate of speech, various F0 measures, and age. He found that speaker recognition was lower for the set of similar voices than for distinct voices. Further, Papcun, Kreiman, and Davis (1989) defined voices based on their recognizability. They termed them easy-to-remember and hard-to-remember voices. A hard-to-remember voice carries less distinctive features than an easy-to-remember voice. Papcun et al. based their analysis of voice memorability on perceptual evaluation and decline in listener recall ability.

4.1.6 Disguise
A factor that has impact, and greatly so (Doherty & Hollien, 1978), on speaker identification is the use of disguise. A disguise can be anything from whispered speech, talking with a raised or lowered F0, dialect mimicry, foreign accent imitation, change of speech rate, and using an artificially induced creaky voice (Künzel, 2000; Masthoff, 1996). These all can be made without external manipulation of the voice. The effect of these disguises vary, where some can even make the speech unintelligible but mostly the goal is to alter the voice enough to make an identification impossible or difficult.

4.1.7 Emotions
Emotion as a factor for speaker identification has received little attention. Read and Craik (1995) recorded actors reading emotional and non-emotional statements and presented listeners recordings of these. They found that the level of emotional content did not impact to any greater extent on listeners’ ability to recognize the speakers than more neutral recordings. However, the acoustic features that are related to emotional utterances have been extensively investigated (e.g. Scherer, 2003; Schröder, 2004). These features often overlap with those found to be prominent in speaker recognition and speaker discrimination, which, in turn, makes emotions in speech a difficult property to deal with in speaker identification processes.

4.2 Measureable factors
The previous section presented factors that need to be interpreted from an acoustic signal to be classified correctly. This section presents factors that are directly measureable in the acoustic signal. However, whether
all these factors are used by listeners during speaker recognition is not known.

4.2.1 Formant transitions

The previously presented factors in (section 4.1.) relate to the indices presented by Abercrombie (1967). However, features that are more easily available in the acoustic signal have also been investigated for their usefulness in speaker identification. Such factors are formant values (e.g. Brown, 1981; Hollien, 2002) and formant transitions (e.g. Greisbach, Esser, & Weinstock, 1995; Ingram, Prandolini, & Ong, 1996; McDougall, 2004, 2005; Rose, 1999). One of the rationals behind measuring formant values over time is that it is the movement between target sounds that carry more individual differences than the target sounds themselves. As Nolan (2002) put it:

Most of our acoustic-phonetic knowledge, and most of our formant-related characterization of speakers, has an essentially static nature. We concern ourselves for instance with vowel centre frequencies, and [...] ‘loci’ characterizing the point to which a formant moves for a consonant of a given place of articulation. I would suggest that the imprint of an individual’s speech mechanism (language, articulatory habits, and vocal tract anatomy combined) will be found to lie more in dynamic descriptions than in static descriptions. (p. 81)

McDougall (2005) further related the movements in the speech apparatus with those in human movement and argued that since people can be recognized by their gait (e.g. Nixon & Carter, 2006), speakers should carry their individuality in their speech apparatus movements, and thus their formant movements. The movements have been termed differently by different researchers, e.g. formant dynamics (McDougall, 2004), F-patterns (Elliott, 2001; Rose, 1999), formant contours (Greisbach et al., 1995) or formant trajectories (Ingram et al., 1996).

The length of the segment and the spectral diversity of the segment in which the formant values are measured also impact upon the ability of the measures to separate speakers (Greisbach et al., 1995; Ingram et al., 1996). Greisbach et al. found that several measure points across time were significantly better than using single measure points, such as the midpoint of a single vowel. Further, they found that using spectrally more diverse segments (i.e. diphthongs) increases the separability of the speakers. Ingram et al. used longer segments to investigate the same effect and found that longer segments performed better. Further, Ingram et al. found that weak segments (often associated with connected speech processes (such as schwa) contained little individual speaker information in the formant transitions.
Rose (1999) investigated the difference between recordings of single speakers made at different times. For this purpose he used recordings of the short utterance Hello, in Australian English. Rose used time alignment by setting measure points at acoustic events; in total seven measure points for each of the first three formants. He measured the difference for each measure point and found that the within-speaker variance was lower over time than the between-speaker variance. That is, each speaker differed less in their production of the segment, as represented by the individual measure points, over a time period of up to four years, than each speaker’s production compared to other speakers’ productions.

Rose (1999) used a single word, experimentally controlled and produced several times, to separate speakers. Albeit a common word, it may be difficult to find such similar segments in realistic settings. Rodman, McAllister, Bitzer, Cepeda, and Abbitt (2002) argued that using ‘isolexemes’ remedies that problem. Isolexemes are segments of sounds that stem from similar words or segments. That is, isolexemes “may consist of a single phone . . . ; several phones such as the rime . . . of a syllable . . . ; a whole syllable: a word; sounds that span syllables or words; etc.” (p. 26) In effect the selected ‘isolexeme’ may be of arbitrary length but should capture individual speaker differences. However, see E. J. Eriksson, Cepeda, Rodman, McAllister, et al. (2004) for a discussion on isolexemic length in the method applied by Rodman et al. (2002).

One isolexeme was investigated by McDougall (2004). She used formant dynamics to study the effect of speaker variance on the production of the Australian English diphthong /aI/. She recorded five Australian English speakers in a laboratory and manipulated both speech rate and prosodic stress. McDougall took formant measures at equidistant points across time throughout the diphthong. She then used these measurements to investigate the usefulness of the points as predictors of speaker identity. As tool of analysis she used linear discriminant analysis (LDA) and found that, at a level of 95%, correct classification could be achieved with these measurements and the LDA technique using a cross-validation method. She did not find an impact of speech rate but concluded that “the nuclear- and non-nuclear-stressed tokens should be compared separately” (p. 124).

One of the drawbacks of using formant transitions as a source of speaker identity is that they are susceptible to different speech processes. For instance, F2 movement has been shown to vary due to speech rate (Tjaden & Weismer, 1998), and lenition and co-articulation affect the length of the transition and how it is displayed (Ingram et al., 1996). See Strange (1987) for a review of the information contained in formant transitions.
4.2.2 Fundamental Frequency
A feature that has been found to correlate with speaker identity is the fundamental frequency (F0) (van Dommelen, 1990; Gelfer, 1993; Walden et al., 1978; Wolf, 1970). The feature is, however, also correlated with other, more general, descriptions, such as regional dialect (e.g. G. Bruce, 1998) and emotions (e.g. Schröder, 2004) which makes it difficult to draw any conclusions about this specific feature. Further, fundamental frequency was concluded to be highly salient in gender discrimination (Lass et al., 1976).

4.2.3 LTAS
The long term average spectrum (LTAS) is a description of the spectral content of a segment measured (Pittam & Rintel, 1996). It has been argued to be effective in speaker discrimination processes (Doherty & Hollien, 1978; Hollien & Majewski, 1977; Hollien, 2002; Kiukaanniemi, Siponen, & Mattila, 1982). It has, however, also been argued to display voice quality differences (Hollien, 2002; Tanner, Roy, Ash, & Buder, 2005), been used to successfully differentiate between genders (Mendoza, Valencia, Muñoz, & Trujillo, 1996), and has been found to display talker ethnicity (Pittam & Rintel, 1996). LTAS is computed by calculating consecutive spectra across the chosen segment and then taking the average of each frequency interval of the spectra. However, it may be unstable for short segments (Pittam & Rintel, 1996).

4.3 External factors
Not only acoustic and perceptual factors that are carried by the voice influence listeners’ ability to judge speaker identity. External factors such as acoustic environment and contextual cues may impact on both the listeners accuracy in recognizing speakers (e.g. Ladefoged, 1978; Kerstholt et al., 2006; Zetterholm, Sullivan, & van Doorn, 2002) and their confidence of making the correct identification (Kerstholt, Jansen, van Amelsvoort, & Broeders, 2004; Olsson, 2000; Yarmey et al., 1994).

4.3.1 Retention interval
Some researchers have reported degradation of recognition after periods of time (Kerstholt et al., 2006) and for certain kinds of voices (Papcun et al., 1989). However, Saslove and Yarmey (1980) found no reduction in recall rates after 24 hours compared to immediately following point of encoding but both Kerstholt et al. (2004) and Kerstholt et al. found reliable degeneration in recognition accuracy after a week, but after three and eight weeks the difference in recall levelled off. Papcun et al. (1989) also investigated the impact of retention intervals and found that listeners’ ability to recognize speakers decreases over time; they also found that this ability is affected by the voice’s qualities; its distinctiveness.
4.3.2 Sample duration and quality
Read and Craik (1995) tested a range of variables and their respective impact on speaker recognition. Two of these variables were the content and the amount of the material presented. Read and Craik found that listeners were unable to identify a speaker by voice alone if the statement length during testing was brief (approximately four seconds) and the tone of which it was uttered changed from conversational to emotional. By increasing the similarities between the contents of test and training material and the way these two are uttered, the accuracy of which speakers are recognized increased. However, Yarmey (2001) found that the content of the utterance did not correlate with listeners’ accuracy in speaker identification if longer passages of training material were available to the listeners. Similarly, Cook and Wilding (1997a) and Roebuck and Wilding (1993) found that recognition accuracy of speakers increased with sample length (used for training) but did not increase with segment (vowel) variety. Pollack, Pickett, and Sumby (1954) found a non-linear relationship with speech sample length such that with samples shorter than a monosyllabic word “speaker identification was only fair” (p. 404). On the other hand, Compton (1963) found that familiar speakers can be accurately identified from as little as 1/40th of a second, if content is kept fixed (a stable vowel).

4.3.3 Speaker familiarity
Yarmey, Yarmey, Yarmey, and Parliament (2001) found effects of familiarity with the target voice in that highly familiar voices were recognized faster and more accurately than less familiar voices. As described in section 4.3.3, the longer the training material the better the recognition accuracy, but Yarmey et al. argued that for highly familiar voices the length effect is only marginal since the identification rates are high from the beginning. Further, Read and Craik (1995) found that the familiarity of the target voice had no impact on recognition if the speaker was left unidentified during training. That is, if listeners fail to recognize (i.e. name) the speaker during the encoding phase, they have no benefit of their prior familiarization.

In order to for a speaker to become familiar exposure to the speaker is necessary. Cook and Wilding (1997b) had listeners familiarize themselves with speakers presented with sentence length samples. However, when Cook and Wilding tried to compare the results of their experiment with a model for familiar face recognition (V. Bruce & Young, 1986) they came to the conclusion that the speakers in their sample set were not familiar to the listeners. They further argued that such a short sample length (one sentence) may not be enough to make a speaker familiar to a listener.
Speaker familiarity was also found to have an impact upon listeners’ ability to shadow voices but only when the speaker was identified (i.e. named) (Newman & Evers, 2007). If the voice to shadow was known (both identified and familiar) listeners were significantly better at attending to that voice than when trying to attend to unfamiliar voices.

In speaker similarity judgements, Walden et al. (1978) found no effect of speaker familiarity. That is, listeners did not use any other perceptual space when analysing familiar speakers than when analysing unfamiliar speakers.

4.4 Factor summary

In sum, a range of factors have been correlated or found to be important in speaker recognition. These are all related to the original set of indices that Abercrombie (1967) defined. The features presented include the speaker’s gender, age, and regional or foreign accent. In addition, other factors not related to the voice production impact upon the listeners’ ability to detect speaker identity. These include retention interval, sample duration and speaker familiarity. Further, acoustic features that are immediately available from the voice signal can be used to separate speakers. These include LTAS, fundamental frequency and formant transitions.

How the features interact and their individual saliency is currently not completely mapped. It has been proposed that there is not a fixed set of features identifying each individual speaker, instead each speaker is delimited by a set of features; which features the set is constructed of varies between speakers (D. Van Lancker, Kreiman, & Emmorey, 1985). The same argument was asserted by van Dommelen (1990, p. 259):

> the relevance of perceptual cues in the recognition of familiar voices was shown to be not hierarchically fixed, but depend on speaker-specific voice characteristics

The results of speaker similarity judgement studies have been inconclusive (see Gelfer, 1993; Walden et al., 1978; Murry & Singh, 1980). The lack of conclusive results was predicted by D. Van Lancker, Kreiman, and Emmorey (1985). If different speakers are defined by different feature sets then correlating psychological dimensions with targeted features will prove useless since these dimensions will correlate with different features depending on the speaker.
5. Materials and Papers

This section presents the database that was used in E. J. Eriksson and Sullivan, (n.d.-b; Paper 5) and E. J. Eriksson, Cepeda, Rodman, McAllister, et al., (2004; Paper 6) and a summary of the Papers included in Part II of this Thesis.

5.1 UDID – Umeå disguise and imitation database

The database used as source for E. J. Eriksson and Sullivan, (n.d.-b; Paper 5) and E. J. Eriksson, Cepeda, Rodman, McAllister, et al., (2004; Paper 6) is the Umeå Disguise and Imitation Database (UDID) that was set up as part of the project Imitated Voices: a research project with applications to security and the law funded by The Bank of Sweden Tercentenary Fund (Dnr: K2002-1121:1–4).

The database consists of recordings of 29 speakers, 17 males and 12 females made in a sound attenuated room. Each speaker was asked to read a newspaper text followed by an interview about the text with a recording assistant. The newspaper text was handed to the participants one week prior to the recording session and they were all asked to familiarize themselves with the text so that they could read it as fluently as possible. Each reading took approximately 3.5 minutes and the following interview lasted about 15 minutes. Thus, both read and spontaneous speech was recorded from each speaker. In addition, two more recordings were made by each speaker. First they were asked to scream, as loudly as possible, a short excerpt from the text read (a single sentence) and then read the same excerpt with a loudness between talking normally and screaming (each speaker made their own subjective evaluation of the loudness chosen for this recording). During the screaming and talking loudly recordings the speakers were asked to face away from the microphone to reduce ceiling effects and clipping by the microphone or recording equipment. Further, these last two recordings were repeated until a successful recording without clippings, misreadings or other artefacts was completed. All speakers in the database received a cinema ticket cheque after completion of their recordings.

The recordings were made onto either a DAT recorder, or a combination of a DAT recorder and a personal computer. If the material was recorded with a DAT recorder, it was later transferred to a personal computer. The material was initially digitized at 48000 Hz, but later down-sampled to 16000 Hz on a personal computer. Further, the material was high-pass filtered at 60 Hz. The material on the DAT tapes were left untouched as reference material.

The spontaneous speech material was interspersed with the interviewer’s voice and with overlaps between the two participants. These
files were labelled and cut to remove the interviewer’s voice and overlaps. The overlaps were kept, however, and labelled appropriately.

Part of the project concerned amateur voice imitations. Therefore, of the 29 speakers recorded, three male speakers were selected as imitation targets. They were selected based on their dialectal background (one from the south, one from the north, and one from the central part of Sweden). Six males previously recorded for the database, were asked to imitate the three target voices. The imitators were also selected on basis of their dialectal background (two from the south, two from the north, and two from the central part of Sweden). Of these six imitators, only five completed all three imitations (one from the northern part of Sweden dropped out after one imitation). These imitators had no, or very little, prior experience with voice imitation.

The amateur imitators were given training material approximately one week prior to recording. This material consisted of a CD of the target voice reading the newspaper text. They were given one target voice at a time and were not given the next target voice until they had been recorded imitating the previous voice. This protocol was designed to minimize the imitators’ confusion between target voices. The imitators were further asked to keep a diary of how much they trained. They trained, on average, about 4 hours per voice spread across the week. Additionally, they were given no further instructions and could approach the imitation task in any way they chose.

The imitation recordings were made in the same way as in the original recordings: first they read the newspaper text (using the imitation), then they were asked to discuss the text, still using the imitation. Finally, they were asked to scream and talk loudly still imitating the target voice. This procedure was repeated for all three target voices. This way, imitations were collected for both reading and spontaneous speech, and screaming and talking loudly. Regardless of the success of the imitations no one was asked to repeat their imitation.

One year after the original recording the five imitators were asked to re-read the newspaper text with their own voice to provide non-contemporaneous speech material. Again, the participants were asked to read the newspaper text and to scream and talk loudly the sentence chosen a year before.

All recordings were labelled according to their content; whether it was read or spontaneous material, whether it was imitated material, whether it was a sentence read screaming or talking loudly and whether it was collected one year after the initial recording. Thus, the structure of the database is based on speaker id (encoded with gender), type of material (read or spontaneous) and type of content (original, one-year delay, imitation, and screaming or talking loudly).
5.2 Summary of Papers

The following section briefly summarizes the Papers included in Part II of this Thesis.

5.2.1 Paper 1 – Emotions in Speech: Judicial Implications

This paper includes a review of the perceptual and acoustic research on detection of emotions in speech and a presentation of the variables in the speech signal found to be affected by emotions. Further, implications for forensic and judicial areas are detailed, these include witness reliability when affected by emotions, both at encoding time and at point of recall, and descriptive functions of emotions for the judicial system.

5.2.2 Paper 2 – Acoustic Impact on Decoding of Semantic Emotion

This paper presents an experiment on how semantic emotions affect the decoding of acoustic emotions carried in the signal. Listeners were presented with recordings of emotional utterances. These utterances carried an acoustic and a semantic emotion. The acoustic emotions could either match (congruent material) or not match (incongruent material) the semantic emotions. Listeners were instructed to answer either yes or no to a question “is this speaker X?” where X could be angry, happy, sad or neutral. The questions could be answered either in relation to the acoustic or semantic emotion carried. Further, sometimes the question emotion matched neither of the acoustic or semantic dimensions.

The results showed that the listeners’ decoding of emotionality in congruent recordings was well above chance. That is, listeners were able to decode the intended emotional (semantic and acoustic) content in recorded sentences that matched the two modalities. Further, listeners were still able to identify the emotion carried, when presented with incongruent recordings. The recognition of semantically carried emotions, however, were significantly lower than for the acoustic emotions.

Analysis of the reaction times for the congruent and the incongruent recordings showed that listeners were significantly slower when responding to incongruent material. Further, when analysing correct responses only listeners were significantly slower in responding to the questions which involved the semantic content compared to questions about the acoustic content.

In sum, without imposing limitations on listeners’ choice of emotional representation the study showed that listeners decode acoustic emotions more rapidly than semantic emotions. Further, the only effect a mismatch between semantic and acoustic emotion impose is an increase in response time; accuracy is not affected.
5.2.3 Paper 3 – On the perceptual dominance of dialect
The impact of the regional dialect as a feature in speaker recognition was investigated. In order to keep other factors fixed a bidialectal speaker who reported using two regionally disparate dialects was used in a voice line-up test. The findings were fourfold. First, listeners could reliably identify the speaker from the line-up when the dialect matched between the training session and the test. Second, listeners failed to recognize the speaker if one dialect was used in the training session and another during the voice line-up. Third, listeners’ familiarity with one of the regional, dialects defined as having grown up and living in the region, did not impact on their ability to select the target voice when the speaker shifted dialects between training and testing. Fourth, and finally, listeners were still unable to decode speaker identity during testing even after being explicitly told that the speaker would be changing the dialect between the training and the test phases. It was concluded that the feature dialect is a powerful and highly salient feature in speaker identification.

5.2.4 Paper 4 – Dialect imitations in speaker identification
Farrús, Eriksson, Sullivan, and Hernando, (in press; Paper 4) investigates the impact of language awareness (background knowledge) on listeners’ ability to detect accent imitation and sensitivity to switched accents. To test this material from a native American English speaking professional actor who had imitated various accents, both English dialects and Spanish accented English was collected from a set of movies. In three experiments listeners were asked to judge (i) whether two samples played were from the same speaker, (ii) whether two voices were spoken with same accent, and (iii) the origin of the speaker from a closed set of regions. Results indicated that the quality of the accent imitation is correlated with listeners’ ability to judge whether voices come from the same speaker. That is, listeners were unable to assign the same speaker to the presented two samples if the quality of the accents presented is good. Further, the listeners’ background information about the accents in the tests predicts their susceptibility to accent imitation. That is, listeners with high degrees of familiarity with a specific accent are less likely to be fooled by an imitation.

5.2.5 Paper 5 – An investigation of the effectiveness of a Swedish glide + vowel segment for speaker discrimination
The paper presents an investigation of whether a method developed using Australian English diphthongs to separate speakers could be applied to a Swedish glide + vowel transition. Swedish contains fewer diphthongs and thus another segment of similar kind had to be selected. Further, the Australian English study used highly controlled test material, whereas the Swedish material contained less control in the produc-
tion of the target segments. Five native speakers of Swedish provided readings of a newspaper text which included seven repetitions of the glide + vowel transition /jœ/. From these segments formant transitions measurements were collected.

The method used in the Australian English study used linear discriminant analysis followed by a cross-validation method to evaluate the separability between the speakers. The Swedish transition showed comparable results to the Australian English diphthong. These results held even when a more stringent cross-validation method was applied and when the sample sets collected were smaller and contained fewer controlled segments. It was also found that the more features included in the analysis the better the separability of speakers.

5.2.6 Paper 6 – Cross-language speaker identification using spectral moments

This paper outlines a short experiment that investigated the possibility of extending a method for separating speakers based on their spectral representation from an American English setting to a Swedish one. The method utilized the mean and variance (the first two spectral moments) of a speaker’s spectrum over time to distinguish (by linear discriminant analysis) between speakers. The results indicate that the tool was successful in separating the speakers, even though the material was markedly less dynamic in the Swedish data (stable vowels were used). However, compared to the power of discrimination between the speaker for the American English data, the Swedish stable vowel data were considerably worse.

5.2.7 Paper 7 – Robustness of Spectral Moments: a Study using Voice Imitations

This study was designed to test the robustness of the spectral moment method presented in E. J. Eriksson, Cepeda, Rodman, McAllister, et al., (2004; Paper 6). Two professional and three amateur imitators were employed to imitate a speech given by a Swedish politician. The material used in the test were collected from the imitations, recordings of the imitators original voices (excluding one professional imitator who did not perform this recording), and the recording of the politicians speech. The professional imitators had previously been found to mislead listeners into selecting them as originating from the target of the imitation (Schlichting & Sullivan, 1997; Zetterholm et al., 2002). The material was analysed both by linear discriminant analysis and by Mahalanobis distances between the speakers.

The results show that although listeners previously had been found to misattribute the imitations to the target voice (Schlichting & Sullivan, 1997; Zetterholm et al., 2002), the method of spectral moments did
not. Further, the Mahalanobis distances revealed two things. First, the target speaker (the politician) is not close to any of the imitations (interestingly, it is closer to the original voice of one of the speakers than that participant’s imitation). Second, the professional imitator (the one who provided his original voice) is better at disguising their voice than the amateur imitators – the distances are smaller between the amateurs’ imitations and original voices than the distances between the professional’s imitation and original voice. It was concluded that spectral moments are robust and their power is not reduced by imitations, even professional ones.

5.2.8 Paper 8 – Effects of age and age-related hearing loss on speaker recognition or can senior citizens be reliable earwitnesses

E. J. Eriksson, Czigler, Skagerstrand, and Sullivan, (n.d.; Paper 8) report an investigation of whether age and corrected hearing loss impact on listeners’ ability to identify speakers. The task given to the listeners was to select a target voice from a voice line-up after a training session. The training material was an imitation of a Swedish politician. Two experiments were undertaken by three groups of participants, an adult group in the age range of 21 to 40, a group of elderly listeners aged 60 years and above without a reported hearing problem, and a group of elderly listeners aged 60 years and above with a reported and corrected hearing problem. In experiment one, the groups were trained with a political speech, i.e. expected context of the speaker being imitated and in experiment two the groups were trained on an imitation of a reading of a cake baking recipe, i.e. unexpected content. The results indicated that elderly listeners with and without a reported hearing problem performed as well as adult listeners without a reported hearing problem when the content was expected. However, when the content was unexpected (i.e. the cake baking instructions) the elderly group performed significantly worse than the adults. That is, elderly listeners relied more on content.

6. MEMORY MODELS OF SPEAKER RECOGNITION

The data presented in Part II can be discussed in relation to models of speaker recognition. Three models form the basis for the discussion in this Thesis: D. Van Lancker, Kreiman, and Emmorey (1985), Papcun et al. (1989) and Belin, Fecteau, and Bédard (2004). These three models try to explain how processing of familiar and unfamiliar voices function. The following section contains a presentation of each model and followed by a discussion about how the findings in the papers presented in Part II related to the models.
6.1 Pattern recognition model

Recognition of familiar voices has been repeatedly shown to be efficient. D. Van Lancker, Kreiman, and Emmorey (1985) argued that most pre-1985 research on speaker recognition had a major flaw: the use of unfamiliar voices. They considered recognition of familiar voices to be different from any other type of voice analysis. D. Van Lancker, Kreiman, and Emmorey and D. Van Lancker, Kreiman, and Wickens (1985) used famous speakers to evaluate listeners' accuracy in recognizing famous voices. D. Van Lancker, Kreiman, and Emmorey investigated how playing recordings backwards to remove timing and phonetic information affected listeners' ability to recognize the famous speakers. They found that, although listeners on average performed well, a striking difference was found between speakers. That is, listeners performed well on some voices and less so on others.

D. Van Lancker, Kreiman, and Wickens (1985) also considered how speech rate interact with speaker recognition and again found that listeners performed differently for different familiar speakers. Based on the results from D. Van Lancker, Kreiman, and Emmorey (1985) and D. Van Lancker, Kreiman, and Wickens D. Van Lancker, Kreiman, and Emmorey proposed that “… voice pattern as an acoustic signal contains a constellation of potential cues from which the listener ‘selects’ a subset to use for identifying a give voice (or, put another way, each voice pattern ‘offers’ a unique subset of qualities)” (p. 33) and that a “[l]oss of one parameter will not impair recognizability if a voice is sufficiently distinctive on some other dimension(s)” (p. 33) and hence argued that “it is not useful to pursue the parameter that contributes most universally to voice identity” (p. 33). D. Van Lancker, Kreiman, and Emmorey finally concluded: “… recognizing a familiar voice is essentially a pattern recognition event …” (p. 34)

Results presented in this Thesis affect the prototype model. E. J. Eriksson, Schaeffler, Sjöström, Sullivan, and Zetterholm, (n.d.; Paper 3) and Farrús et al., (in press; Paper 4) presented data that show that listeners' ability to detect a speaker's identity is affected by the content carried in particular features. E. J. Eriksson, Schaeffler, et al. (Paper 3) showed that listeners are unable to identify a speaker when the dialect spoken at point of learning is different than at point of recall. Further, Farrús et al. (Paper 4) showed that listeners are unable to detect that a speaker is the same when the speaker uses two different accents.

In order to incorporate the findings of E. J. Eriksson, Schaeffler, et al., (n.d.; Paper 3) and Farrús et al., (in press; Paper 4) into the model by D. Van Lancker, Kreiman, and Emmorey (1985) the model needs a hierarchy such that some features take presidency over other. The results of E. J. Eriksson, Schaeffler, et al. and Farrús et al. showed that the feature regional dialect should be hierarchically high in the model. Further, the
results of Walden et al. (1978) and Goldinger (1996) for instance, show that age and gender should be at least as high in the hierarchy of the model.

In as far as it is possible to compare mathematical methods with perceptual evaluations, the method used in E. J. Eriksson, Cepeda, Rodman, McAllister, et al., (2004; Paper 6) and E. J. Eriksson, Cepeda, Rodman, Sullivan, et al., (2004; Paper 7) show greater resistance to variation within a speakers voice than human listeners show. The method is based on the two first spectral moments (mean and variance about the mean of the spectrum) taken over time. These values form the basis of the subsequent analysis. However, as shown the spectral moments are much more tolerant to variations in features that listeners are sensitive to. Thus, in the D. Van Lancker, Kreiman, and Emmorey (1985) model, these two values are not part of the feature set used by listeners.

6.2 Neurological model of speaker recognition

Using data from neuro-imaging and psychological studies Belin et al. (2004) adopted the face recognition model of V. Bruce and Young (1986) to voice recognition. Belin et al.’s idea was to describe voice recognition and to link the processes involved to those of face recognition based on neurological evidence.

The face recognition model (V. Bruce & Young, 1986) after initial structural analysis of the face the model forks into three key parts: an expression analysis process, a facial speech analysis process and face recognition units. The latter is defined: “Each face recognition unit contains stored structural codes describing one of the faces known to a person” (p. 311). V. Bruce and Young argued that activation of a face recognition unit is dependent on the amount of match between stored structure and structure provided in the percept. The activation of a face recognition unit can be raised or lowered by expectations surrounding the event through activation in person identity nodes. These nodes are part of the semantic memory and are the only parts of the system that can name an individual; there exists only one node per person and that node carries detailed semantic descriptions of that person, including other modal descriptors, such as voice or gait habits.

Belin et al. (2004) adapted this theory of face recognition to voice recognition. They presented a model that, after structural decoding of the speech signal linked three processes of voice information to the corresponding processes presented in the V. Bruce and Young’s (1986) model (see Figure 6.1). The voice recognition model functions as follows: first, voice information passes through low level auditory systems and is passed on to a voice structure analysis. After the structural analyses the signals are forked into three different functions. One, a vo-
Figure 6.1: Belin et al.’s (2004) model of voice recognition. The left hand side, the voice part, is linked to the right hand side, which is based on the model of face recognition by V. Bruce and Young (1986). Reprinted with permission.

cal speech analysis function; two, a vocal affect analysis function; and three, voice recognition units. Finally, the voice recognition units were linked with the person identity nodes. Belin et al. provided data from brain imaging studies that showed brain regions that would be linked to the functional systems in the model, including the possible overlaps between facial and voice recognition systems. For instance, emotion decoding has been investigated and found to have separate brain substrates from perception of speaker familiarity (Imaizumi et al., 1997) and speaker identity has been found to have different brain regions associated with it than verbal content analysis (Kriegstein et al., 2003). Further, behavioural data have shown that affect can reduce the ability of listeners to correctly assign a speaker to a particular dialect area (Williams et al., 1999, but see Read & Craik, 1995, for opposing information when identifying speakers).

The three systems following the fork (the vocal speech analysis, the vocal affect analysis and the voice recognition units) were argued to be functionally dissociated: “It is only at the highest levels of the architecture that representations for one type of information would become independent of sources of variability related to other types of information.” (Belin et al., 2004, p. 131, Box 3)
The model proposed by Belin et al. (2004) is a compelling model but it does not specify how specific features are correlated with voice familiarity (i.e. how the voice recognition units operate). Further, unfamiliar voices, it must be argued, follow the same structural analysis as familiar voices and the process of unfamiliar voices should be connected to the processes of vocal affect and speech analysis. However, how the process of unfamiliar voices fit into this model is left unexplained.

The papers presented in this Thesis that are related to the model by Belin et al. (2004) are E. J. Eriksson, Rodman, and Hubal, (in press; Paper 1) and E. J. Eriksson, Schaeffler, and Sullivan, (in press; Paper 2). These two papers point to extensive overlap between different types of analysis. E. J. Eriksson, Schaeffler, and Sullivan point to a processing cost associated with parallell but mismatching processing; acoustic analysis takes place in the vocal affect process, and semantic analysis takes place in the speech analysis process. When the two processes arrive at different results, the mismatch must be resolved. However, the accuracy of emotion identification is not affected by the increased processing times.

To some extent, E. J. Eriksson and Sullivan, (n.d.-b; Paper 5) could also be said to related to the voice recognition model by Belin et al. (2004). In this paper it was shown that a segment taken out of context could be mathematically used to separate speakers. The segment was represented by formant transitions which are readily available to listeners (Strange, 1987) but are affected by particular speech processes (e.g. Ingram et al., 1996; Tjaden & Weismer, 1998). The speech process related effect of prosodic stress was investigated by McDougall (2004) who argued that the effect is not detrimental to speaker separation but comparisons should be made within prosodic stress type.

6.3 Prototype model of speaker identification

Both D. Van Lancker, Kreiman, and Emmorey (1985) and Belin et al. (2004) designed their models for familiar voices. Papcun et al. (1989), however, investigated the recognizability and the effects of unfamiliar voices on memory decay and accuracy and argued for a prototype model for voice recognition. Prototypes are memory representations defined by particularly representative exemplars of a category or a set of describing attributes (Eysenck & Keane, 1999). Papcun et al. argued for two types of unfamiliar voices, easy-to-remember and hard-to-remember voices. Hard-to-remember voices demand more time to encode but once encoded are more persistent in memory. Easy-to-remember voices carry large amounts of distinct features but these decay in memory fast. Papcun et al. argued that the prototypes in voice memory are made up of hard-to-remember voices. They based their argument on a number of results. First, they found that easy-to-remember
voices decay in memory faster than hard-to-remember voices. Second, they found that hard-to-remember voices are more easily assigned as being the target voice than easy-to-remember voices. Based on these results they concluded that hard-to-remember voices form the prototype of a memory concept and forming new concepts demand greater encoding time.

Papcun et al. (1989) also found that easy-to-remember voices were less frequently falsely assigned to a particular prototype, even after prolonged retention times. The reason for this was argued to be that the complete voice description is available at presentation and since easy-to-remember voices carry highly distinguishable features these can easily be ruled out. This explanation accounted for their test participants’ errors:

When hard-to-remember voices are targets and easy-to-remember voices are probes, there will be relatively few errors because the stable target voice characteristics as well as the immediately present probe voice characteristics are available to the decision. In the converse case, however, an easy-to-remember target voice will lose some of its characteristics. Hence, when a prototypical voice is used as a probe, more errors are to be expected. (p. 923)

In the experiments presented in the E. J. Eriksson, Schaeffler, et al., (n.d.; Paper 3) and Farrús et al., (in press; Paper 4) it is shown that it may not be hard-to-remember voices that are at the bottom of the prototypes. The bottom of these prototypes may consist of only a few descriptors (i.e. not exemplars but a collection of attributes) of which one is the feature gender (Goldinger, 1996), and one is the feature regional dialect (E. J. Eriksson, Schaeffler, et al., n.d.; Paper 3). For instance, a speaker from one dialect cannot be assigned to the correct prototype if the speaker changes dialect or changes his voice in such a way that dialectal features become unstable.

6.4 Discussion

The model of D. Van Lancker, Kreiman, and Emmorey (1985) and of Belin et al. (2004) can be merged. Belin et al.’s description of their processes is limited in detail. On the other hand, D. Van Lancker, Kreiman, and Emmorey ignored the processes prior to familiar voice identification. The model of pattern matching by D. Van Lancker, Kreiman, and Emmorey can thus be fitted into the process of voice recognition units in Belin et al.’s model. That means that proper activation of a pattern for an individual gives rise to activation in person identity nodes, which include the name of the individual. However, according the V. Bruce
and Young (1986), the person identity nodes should be able to, due to
texture or expectancy, enhance, or facilitate, recognition for specific
individuals in the voice recognition units. Thus, it may be that the per-
son identity nodes could ‘fill in the gaps’ for known voices which would
lead to greater tolerance for variation within certain features.

As has been showed in the Papers presented in this Thesis, some fea-
tures appear to be hard-coded into the person identity nodes (such as the
feature dialect) which either leads to a failure of activation of the proper
person identity nodes (bottom-up effect) or prior activation among the
wrong person identity nodes inhibit activation of the right voice recog-
nition units (top-down effect). In both cases, naming fails due to the
mismatch between the voice recognition units and the person identity
nodes.

Another way the person identity nodes could interact with the pat-
tern recognition model proposed by D. Van Lancker, Kreiman, and Em-
morey (1985) is that the nodes fill in which features are relevant to iden-
tify a specific speaker including the expected value (or ranges of allowed
values). Some features may carry large ranges, whereas others may only
be defined by a fixed value, For instance, large variation may be ac-
cepted in speech rate, intensity or F2 movement, but no variation at all
is allowed in the feature dialect.

Expectation activates specific person identity nodes, which in turn
activates specific voice recognition units. That way, a person may be
identified faster, or more accurately, if the context is right. Further, if the
expectations mismatch the content of the percept (as in Zetterholm et
al., 2002 and E. J. Eriksson, Czigler, et al., n.d.; Paper 8) then the person
identity nodes would activate the wrong voice recognition units: a voice
would be misrecognized.

The person identity nodes proposed to be connected to the voice
recognition units (Belin et al., 2004) could also be linked to the other
processes. This means that linguistic content may raise the level of acti-
vation in a person identity node. Similarly, a specific affection signalled
by a voice may do the same thing. The other way around may also
function. That is, a specific speaker is identified and thus leads to easier
recognition of what is said due to enhanced robustness to variations in
certain language traits that the identified speaker uses.

So far, the discussion has been focused on the idea that the target
voice is made familiar to the participants in each of the applicable tests.
However, as Cook and Wilding (1997b) pointed out they failed to match
their experimental data to the model proposed by V. Bruce and Young
(1986) and argued that it may be due to insufficient training material.
That is, they argued that the amount of training material presented to
participants was too little (too short) to make the voice familiar. It was
judged as a previously heard but unfamiliar voice. The same argument
may be applied here. The voices heard during the test in E. J. Eriksson, Schaeffler, et al., (n.d.; Paper 3) could be seen as unfamiliar, but previously heard, voices. The comparison to D. Van Lancker, Kreiman, and Emmorey (1985) is then not applicable, based on the double dissociation of the processes for recognizing familiar voices and analysing unfamiliar voices (D. Van Lancker & Kreiman, 1987; D. R. Van Lancker et al., 1988). Thus a model that is designed for unfamiliar voices may be more appropriate.

The final step is fitting the process of analysing unfamiliar voices into the model presented by Belin et al. (2004). The processing of familiar voices and unfamiliar voices was shown to be different (D. Van Lancker & Kreiman, 1987; D. R. Van Lancker et al., 1988). However, the model by Belin et al. contains general speech processing and affects processing as separate functions, in addition to, the voice recognition units for familiar voices. The process of unfamiliar voices needs to be fitted somewhere and it difficult to argue that unfamiliar voices are not submitted to the low-level auditory process and voice structural analysis presented by Belin et al. Thus, unfamiliar voices must be analysed as a separate process at the same level in the hierarchy as the voice recognition units (process of familiar voices). Further, interaction between the process of unfamiliar voices, vocal affect and speech analysis can exist. Perhaps there is also an interaction between familiar and unfamiliar voices. For instance, person identity nodes can be activated for a specific voice selecting specific features that is prepared in the voice recognition units. However, these fail to reach identification and the analysis is turn over to the process of unfamiliar voices.

The process of recognizing unfamiliar voices may work as proposed with prototypes of specific features at the bottom. The process of recognizing unfamiliar voices may also be connected to the person identity nodes in such a way that a similar sounding voice, but previously known, receive a slight increase in activation and the identity is then taken from that node instead of arriving at identification failure. Thus, expectation and attention which work differently in their access of person identity nodes have a much greater impact on the process of recognizing unfamiliar voices.

7. Conclusion
Nine experiments were conducted that identified features that can be used to separate and identify speakers by their voice alone. One finding in this Thesis is that the feature regional dialect is part of a set of critical features used by listeners to identify a speaker by voice alone. A feature akin to dialect is foreign accent which was shown to be impacted by listeners’ prior knowledge of the accent. Further, the formant dynamics
of a transition of a glide + vowel in Swedish may be used with the same level of results as that found for a diphthong in Australian English.

The validity of these features in relation to models of speaker recognition was discussed. The models applied to different parts of speaker recognition processes and each was addressed in turn. The main finding is that the process proposed by D. Van Lancker, Kreiman, and Emmorey (1985), that familiar speaker identification is a process of pattern matching, is the most likely, albeit equally vague in its definition. Papcun et al. (1989) addressed the problem of unfamiliar speaker recognition and applied a theory of prototype memory organization. Papcun et al. suggested that each prototype is defined by an exemplar of a hard-to-remember voice is not supported by the findings in this Thesis. Instead it is more likely that each prototype is defined by a set of attributes, of which gender and dialect are prominent. The model of voice recognition of Belin et al. (2004) was built upon the model of familiar face recognition by V. Bruce and Young (1986). The model proposed by Belin et al. (2004) can be extended to support the findings of the Papers summarized in this Thesis. A larger amount of inter-connectivity between the forked processes in the auditory speech analysis and an inclusion of a process to handle unfamiliar voices should be made.

The findings presented in this Thesis, in particular the findings of the individual papers in Part II, have implications for criminal cases in which speaker recognition forms a part. The findings feed directly into the growing body of forensic phonetic and forensic linguistic research.

8. SUGGESTED AREAS FOR FUTURE RESEARCH

The research presented in this Thesis can be developed in a number of ways. Three are briefly outlined here. One, as has been shown, external factors impact upon listeners’ ability to identify a speaker (see section 4.3). However, the factors discussed in this Thesis can extended with, for instance, noise. The pilot study presented in (E. J. Eriksson & Sullivan, n.d.-a; Paper 9) investigated the impact of uncontrolled noise on dialect recognition. This pilot investigation showed that noise only marginally affects listeners’ ability to identify a spoken dialect from a closed set of alternatives. However, the results in the study need to be contrasted with results of the same test in noise controlled environments to delimit the effect of (or lack of) different noise types at various dB-levels. Further, noise needs to be investigated as a factor in speaker, as opposed to dialect, recognition.

Two, variance have been argued to be important in speech perception (e.g. Mullennix & Pisoni, 1990; Pisoni, 1997; Remez et al., 1997, 2004; Sheffert et al., 2002) as it reinforces memory representations of speech sounds. The presentation of voices representing variation within a di-
alect has been shown to improve dialect recognition (Clopper & Pisoni, 2004b). It is possible that voice familiarity is based on frequency of presentation. If this is the case a familiar voice is familiar only because it is presented a large number of times such that robust representations can be gained. Research to determine when something becomes familiar is needed.

Three, perhaps it is learning that creates a solid representation of specific features that define a speaker as D. Van Lancker, Kreiman, and Emmorey (1985) have argued. They also argued that each speaker can be defined by different sets of features. In this Thesis it is proposed that some of these features have presidency during recognition. It is therefore suggested that familiar voices be investigated in different settings, with respect to their defining features. The aim of such a research programme would be to find the defining features of each voice and then evaluate listeners’ ability to recognize the familiar speakers when presented in a setup of voices with similar defining features.
REFERENCES


