Speaker classification concepts: past, present and future

David R. Hill

Dedicated to the memory of Walter Lawrence and Peter Ladefoged

Abstract

Speaker classification requires a sufficiently accurate functional description of speaker attributes and the resources used in speaking, to be able to produce new utterances mimicking the speaker's current physical, emotional and cognitive state, with the correct dialect, social class markers and speech habits. We lack adequate functional knowledge of why and how speakers produce the utterances they do, as well as adequate theoretical frameworks embodying the kinds of knowledge, resources and intentions they use. Rhythm and intonation - intimately linked in most language - provide a wealth of information relevant to speaker classification. Functional - as opposed to decriptive - models are needed. Segmental cues to speaker category, and markers for categories like fear, uncertainty, urgency, and confidence are largely unresearched. What Eckman and Friesen did for facial expression must be done for verbal expression. This chapter examines some potentially profitable research possibilities in context.

Introduction

Preamble

In an article in the Washington Post published February 1st 1999 science correspondent William Arkin reported on work at the Los Alamos National Laboratory in the US that claimed success in allowing arbitrary voices to be mimicked by computer [1]. One example was a recording purportedly by General Carl Steiner, former Commander-in-chief, U.S. Special Operations Command, in which he said:

> "Gentlemen! We have called you together to inform you that we are going to overthrow the United States government.".

But it was not Steiner. Arkin reports it was the result of voice "morphing" technology developed at the Los Alamos National Laboratory in New Mexico, using just ten minutes of high quality digital recording of Steiner's voice to "clone" the speech patterns followed by the generation of the required speech in a research project led by George Papcun, one time member of the *Computalker* team. Steiner was sufficiently impressed, apparently, that he asked for a copy. Former Secretary of State, Colon Powell, was also mimicked saying something alleged to be equally unlikely.

The process was "near real-time" in 1999, and processor speeds have increased dramatically since then (say two orders of magnitude). This might suggest that the problem of characterising a speaker's voice and using the characterisation to fake arbitrary utterances is so well understood that the speaker classification problem is largely solved, and real-time mimicry no problem. This would be over-optimistic. Of course, the Los Alamos project appears to be secret, and further details are hard to obtain, but a recent job posting in the "foNETiks" newsletter [2]—requiring US Secret security clearance—makes it clear that mimicking human agents in all respects, including even things like eye movements, is an ongoing military research project.

> "Working as part of an interdisciplinary team, this research scientist will help create language-enabled synthetic entities that closely match human behavior. These synthetic entities will be integrated into training simulations to enhance training while reducing resource requirements. These synthetic entities are not gaming systems based on black-box AI techniques. They are cognitively transparent entities that exhibit human-like behavior at a fine-grained level of detail as supported by their implementation in the ACT-R cognitive architecture and validated via psychological measures like eye movements, reaction times and error rates." (From the job description)

Reasons for wanting speaker classification

In approaching the problem of speaker classification, the first question has to be: "What is the purpose of the classification?" because the purpose sets the criteria for the task—a necessary precursor to choosing the tools, techniques and measures of success to be used.

Some of the reasons for attempting speaker classification in one form or another include: (1) a clustering exercise to allow automatic indexing of audio material; (2) identification or verification of individuals for ensuring secure access, including text-dependent and text-independent approaches; (3) determination of speaker characteristics and acoustic environment to facilitate a speech recognition task or tailor a machine dialogue to the needs and situation of the user; (4) a general characterisation of a speaker type to allow a synthetic voice with similar characteristics (accent, gender, age …) to be generated—what we might term "Pronunciation Modelling"; or (5) a very specific characterisation of a particular speaker to allow arbitrary speech to be generated that is indistinguishable from the speech expected from that speaker under a variety of external (physical and acoustic environment) and affective (emotional) conditions—or what we might term "Impersonation". It is interesting to note that, if impersonation could be carried out in real time, it could form the basis of very low bandwidth communication channels, along with other interesting military and commercial applications, including PsyOps. There are other reasons, including forensic analysis for purposes of law enforcement and categorising speakers as part of the methodology in research on dialects and language acquisition. It is in such areas, particularly work by Paul Foulkes (York University, UK) and his co-workers at various universities, that some of the more interesting new directions have emerged in the field of what Foulkes calls sociophonetics [3].

Some reasons for speaker classification, such as identifying the language the speaker is using, can fit into more than one of the categories listed above. The same is true of modelling cognitive and emotional states. In any case, the list is not exhaustive.

More recently, a rising demand for human-like response systems has led to an increasing requirement for the ability to classify speakers in more general ways to permit, for example, machine dialogues to be tailored to clients—allowing shopping systems to recommend suitable goods appropriate to the age and sex of the shopper, or systems that can recognise urgency, confusion or perhaps language impairment. Such systems are reminiscent of the goals of the Air Force Research Laboratory project noted above [2].

When user modeling is important, and interaction is by voice, speaker classification becomes a very broad topic indeed, extending to the environment in which a speaker is operating, as well as the speaker's goals. Paralinguistic and context cues must be extracted along with more traditional speech analysis information and used in the procedures for categorising speakers.

Recent papers providing an entry into the speaker recognition & speaker verification literature include [4]. [5], [6], [7], [8], [9], [10], [11] , [12] and [13]. Two early papers—still highly relevant to speaker identification and verification as well as offering insight into other speaker-dependent attributes of speech—are [14] & [15]. Many of these approaches are reminiscent of those used to try and solve the complementary problem of speech recognition, and typically involve: (a) slicing the speech into consecutive samples of a few milliseconds (the "salami" technique, involving slicing); and (b) carrying out some kind of automated statistical analysis (decision/pattern recognition strategy), on the data collected. Whilst these perhaps pay attention to some knowledge of speech structure—such as formant structure or underlying pitch—they have an unfortunate tendency to take a low-level approach to the treatment of the resulting data rather than an insightful approach (undoubtedly a result of pursuing goals of automating the process whilst minimising the algorithms and CPU cycles involved). The tendency may also result from a necessary de-skilling of the classification task, so that only computing and mathematical skills are needed, even if linguists are consulted. It is tough to create teams, let alone find individuals, who possess all the skills that should be applied to an integrated and informed approach. Given the volumes of data involved, pressure to cast the problem into terms that facilitate uniform machine procedures is considerable. In this book we are concerned with speaker classification rather than verification and/or recognition. It is necessary to turn our attention to speech structures, as these must be dealt with explicitly for more general classification tasks.

More structured approaches

If you are carrying out a statistical analysis of dinosaur bones when studying the characteristics of those prehistoric creatures, your results will not relate to any real dinosaur bones—let alone real dinosaurs—if you disregard the differences between bones due to gender, variety, and so on. Equally, if you wish to gather data relevant to classifying speakers, for whatever reason, you need to understand the attributes of speakers relevant to your required classification, rather than simply hoping that a genetic algorithm, neural net, Gaussian Mixture Model, or whatever will do the job for you. It might, but then again, it very well might not—at least, the classification will be nothing like as good as a properly informed discrimination that takes account of what you know about the populations of interest. This is why I am impressed by Paul Foulkes' work at York. He is doing for speaker classification what I have always regarded as the proper approach for speech recognition., namely carrying out careful experiments to reveal relevant speech structure.

By way of a very simple illustration of what I am talking about, consider what is perhaps the earliest recorded example of a solution to, and application of, the speaker classification problem—as recorded in the Bible:

> The Gileadites seized the fords of the Jordan and held them against Ephraim. When any Ephraimite who had escaped begged leave to cross, the men of Gilead asked him, 'Are you an Ephraimite?', and if he said, 'No', they would retort, 'Say Shibboleth.' He would say 'Sibboleth', and because he could not pronounce the word properly, they seized him and killed him at the fords of the Jordan. At that time forty-two thousand men of Ephraim lost their lives. (Judges 12, verses 5-6 [16])

One has to wonder how the error rate in this classification task compared to the error rate that might have been achieved with modern equipment doing statistical analyses of spectral sections and using Gaussian Mixture Models or similar techniques, but this early classification used very specific information about the population of interest, however flawed the science.

It is essential to use tests that are better than what amount to arbitrary statistical descriptions in order to begin improving our speaker classification, verification and identification techniques and success rates. Much of what is done these days seems little better than recognising faces by comparing captured and stored photographic images, with some kind of statistical analysis of pixel groupings. Some success may be expected, but the approach is inherently limited unless a knowledge of the structure of faces, and the way the substructures vary, and relate to recognition/classification targets, is used.

This is the important step that Paul Foulkes has taken in the speech context, and is the issue to be addressed—in a rather specialised context—in the rest of this chapter. A full treatment is both impossible (as the research has yet to be done), and is beyond any reasonable scope.

Some comments on speaker classification in the context of verification/identification

People vary widely in their ability to recognise speakers, even those speakers they know well. Surprisingly, phoneticians seem no better than untrained listeners in distinguishing between different speakers ([17], quoted by [18]) so it is an open question as to just how the task is being performed. How do listeners who perform well recognise an idiosynchratic accent and idiolect, as opposed to making an accurate phonetic transcription of it? What other subtle cues do listeners use when identifying and distinguishing speakers, and gauging their affective and cognitive state, age and so on that are necessarily unrepresented in our phonetic/phonological models? These are important questions that are worth answering in order to make progress in improving machine procedures and performance at this difficult task. If accurate dialogue models can be constructed to include the use of pauses in turn taking, rhythm, and changes in pitch levels, intonation resources and the like, the give and take in dialogue may offer important clues for speaker classification that are outside current descriptive frameworks aimed at the phonological and semantic record. Research in such areas has been ongoing, though with a view to understanding dialogue behaviour and revisiting meetings rather than classifying the participants—for example, the M4 project [19], as well as [20] and [21].

It is not even clear what categories of speaker and speaker state we could observe, if we knew

how to, let alone what characteristics relate to these categories. We come back to the question of what categories are important, and how can we distinguish them—and thence to the question: "Why do you want to categorise the speakers, and what error rates are acceptable." Presumably those wishing to pass by the Gileadites would have had rather strong views on that topic—especially if they had a lisp!

In their experiment on open set speaker identification by human listeners, Foulkes & Barron [18] found that voices which were less well identified nevertheless contained phonetic cues which were not found in some or even any of the other samples. This suggests that some cues, however salient they may appear to phoneticians, are not particularly useful diagnostics in the process of "live" speaker recognition. This suggests that if the "right" phonetic knowledge is used in structuring the cue determination for automatic speaker recognition it could be even more successful than recognition by listeners who know the speakers well.

In this experiment, even the apparently obvious and well documented clue of "up talk" (high rising terminal intonation, or HRT) was not properly utilised by the listeners attempting to identify their friends, even while they made comments showing that they paid attention to pitch variation. It is interesting that, in pursuing this, the authors carried out a statistical analysis of pitch variation and tied the mean and standard deviation to the results as a basis for their interpretation, rather than examining the intonation patterns in more detail. Part of the reason for this is that we still do not have adequate models for the use of intonational resources. This is one area with which our own research has been concerned, but the research is based on a moving target. For example, Halliday's model of British English intonation (and the associated underlying framework for rhythm) does not include the use of HRT in modern terms, even though the model includes it as a basic option for questions. This introduces some of the questions that have plagued us as we attempted to create a high-quality text-to-speech system—questions that lead naturally into some of the topics I feel are worth addressing. For example, if we had a good functional description of how people use intonational and rhythmic resources to achieve their goals, and a good way of recognising these goals (rather than a description of specific patterns for a particular accent, such as British RP English), such information could likely be used to compare the meaningful differences between different speakers in different situations, and thereby effect a useful categorisation. There is a great deal more choice in the use of intonational and rhythmic resources—at least for speakers without special training—than there is for voice quality, long-term spectral features, vowel quality and the like. Such features, properly extracted, and based on a well understood structure, would be valuable in some forms of speaker categorisation. What we don't want to do is collect unstructured statistics in the hope that something will "pop out" of the data. That way we would simply wind up with a pile of "mythical dinosaur bones"!

As noted above, this chapter is not intended as a survey of speaker classification techniques, but as an outline of problems, possible solutions, and suggested directions for new research.

Some comments on the Foulkes & Barron [18] experiment

In describing their experimental design, Foulkes & Barron write:

> "Like McClelland [22], our study assesses SR by a group of people who know each other very well. Our group, however, consists of a set of young men who are university friends. This group was selected to investigate SR in a situation where the social profile of all group members is very similar in terms of age and gender (compare with McClelland's study, which involved men and women of various ages). …
> "All were male, aged twenty or twenty-one, and formed a close social network. During their first year as students the ten had all lived together in shared student accommodation. Some of the network members spent large proportions of their academic time together, and they had all socialized with each other on a regular basis."

It is well known that one sure way to acquire the "right" accent in Britain is to attend the "right" school. That is how I acquired my own archaic RP accent. I was always amused when—on sabbatical—I crossed the Atlantic (both ways) in the Polish ocean liner "Stefan Batory" in company of other varied academics, diplomats and the like. I came across a family from the United States whose son, after one year at a British public school (British "public schools" are actually exclusive private boarding

schools), had acquired an impeccable RP accent—indistinguishable from my own. His parents were quite embarassed and puzzled. Peer group pressure—especially in a strange environment where the threats are unknown but often quite real—creates a tremendous and largely unconscious pressure to conform in all possible ways, including manner of speech, as I know from personal experience as well as observation. Note that all kinds of subtle cues are assimilated by the newcomer who adapts, and these combine to create a new accent which is acceptable to the group. This undoubtedly makes it more difficult to distinguish individuals.

The Foulkes and Barron experiment was well and insightfully designed, as experiments must be, to maximise the chance of revealing information relevant to the prior hypotheses. If you wish to maximise the chances of confusion, and eliminate the possibility that lack of familiarity contaminates the results with unkown factors, choosing a tightly knit social group—the members of which have actually live together for a year in new surroundings and have acquired similar speaking habits whilst learning to ignore many of the differences—is exactly the right thing to do. You thereby gather useful information that might otherwise have been lost amongst the many possible confounding factors. Note that the chances of confusion were further increased by using telephone speech. In the context of forensics, and psychophysical experimentation, this is entirely appropriate. Working with a group of young people actually at a "public school" might have been even better but perhaps less practical.

Under such circumstances, the perceptually obvious cues that might identify individuals within the group will be considerably attenuated as the members aim to keep a low profile and fit in with the group. The differences that persist, however obvious to a linguist or a machine, are likely to be exactly those cues that are less important for identifying the speakers as different amongst themselves. At the same time, cues which are not so perceptually salient may well still be useful in categorising or recognising speakers by existing machine strategies. More than one approach is appropriate.

In speech recognition and synthesis, the arbiter of acceptable performance is ultimately the human listener—necessarily subjective, which, in turn, means that the perceptual consequences of any given speech determine whether the synthesis or recognition procedures were effective. Consequently, psychophysical studies of speech perception have provided a great deal of important information for those working in both areas.

In speaker verification and recognition, as well as some classification tasks, the ultimate arbiter of success is objective accuracy. Perceptual experiments have received less attention most likely because the classification procedures are amenable to objective measurement—for example, in the case of speaker identification, including foil rejection, classification by age, and classification by gender.

However, if the object is classification by some other criteria, such as accent, emotional state, cognitive state, and environmental effects, the problem once again becomes more subjective since objective measures of success are simply unavailable and success, or lack of it, must once again be based on perceptual judgements, whether by speaker or listener.

The same consideration also applies to a question like: "Does the speaker belong to the group who lived and studied together at university" rather than which particular individual is speaking—the kind of question that may be of considerable forensic interest these days and the answer may not be readily obtained by objective means until after the fact. Foulkes and Barron's [18] experiment shows this very clearly since there were quite salient differences between the speakers in their experimental group that were apparently ignored by human listeners who must therefore have categorised the different speakers using only perceptually relevant cues that somehow had converged considerably towards a uniform state that caused significant confusion even amongst the in-group itself. The features needed to decide that a person was a member of the group were clearly different from the features needed to identify the same person within the group. Perceptual studies may help throw some light on the differences between these features, in concert with other approaches.

Perceptual studies are of interest in their own right, simply as a way of exploring the cues that arise from various factors, such as age, sexual orientation, or in-group membership.

Perceptual studies

In research on recognition and synthesis, perceptual experiments have been powerful tools in resolving the important issues concerned with speech structure.

It might seem, in light of the reports by Shirt [17] & [23] that, since people—even trained linguists—have difficulty recognising speakers, that the perceptual (i.e. *subjective*) effects of differences between speakers are less important than what might be termed the *objective* differences. This would be a misunderstanding as argued in the previous section.

For example, *why* are some listeners able to hear differences that other listeners cannot? Is there a continuous dimension for decision making or is it categorical? If the latter, how does the categorisation threshold vary? Are listeners able to hear differences in the dimensions of interest and, if so what are the difference limens? When there are competing dimensions, which cues are the more powerful? If differences that are theoretically perceptible exist, which are ignored by listeners attempting to recognise speakers, what is the reason? It would seem that conventional linguistic training is not necessarily the issue in these and other investigative questions.

Field work is detailed, painstaking and demanding. Perceptual experiments are no less demanding, even if different skills, methods and tools are required. The work equally requires guidance from all the other subdisciplines of linguistics and psychology. This in itself is demanding but, perhaps, the biggest stumbling block has been the absence of suitable instruments to pursue the experimental work since the cues sought are subtle and not well-understood which therefore demands a high-quality system to generate the experimental material in a controlled, accurate manner, within a matrix that is as natural as possible.

The invention of the sound spectrograph [24] Pattern Playback [25], the Parametric Artificial Talker [26], OVE II [27] and their successors were seminal inventions for our modern understanding of speech structure and speech perception for purposes of speech recognition and speech synthesis.

When the sound spectrograph first appeared, many considered that the problem of speech recognition was close to solution, if not solved, and it was quite a suprise that it took around two years to train observers well enough for them to recognise the "obvious" patterns revealed by the machine. Machine recognition remains a relatively unsolved problem, though programs like *Dragon Dictate* do a reasonably useful job by using a well designed interactive dialogue coupled with training to particular voices. Replicating human abilities is still a dream (and will remain so until we have better ways of representing the real world and using the information effectively—a core AI problem).

Pattern Playback could play back spectrograms of real speech and it wasn't long before Pierre Delattre, at the Haskins Labs in New York, hand-painted such spectrograms based on his experience —with real speech versions and perceptual experiments—to produce the first "real" [sic] synthetic speech. The rules "in his head" were soon made explicit [28], but the speech could not be called natural. It was not long before Holmes and his colleagues produced a completely automated text-to-segmental-level-speech synthesis system [29] to which Mattingly then added automated prosody [30].

With the invention of synthesisers that modelled some of the constraints on the human vocal tract, and Fant's seminal book (based on his thesis presented before the King of Sweden, with Walter Lawrence as the "third opponent") [31], progress in all speech areas accelerated and John Holmes, at the UK government "Joint Speech Research Unit" showed that it was possible to mimic human speech quite closely given enough care in preparing the input data [32]. But this exercise did not formalise just what it was that characterised a particular voice in any way that would be useful for carrying out general categorisation tasks.

Many of the problems that plague solution of speech recognition and synthesis arise from exactly those aspects of speech that reveal information about the speaker. This provides one important reason for wishing to classify speakers, as noted in the introduction. There is also the problem that speech is a continuous acoustic recoding of articulatory gestures possessing no consistently clear boundaries in

the ongoing spectrum of sounds—likened in the early days to an egg that has been scrambled. Both speech recognition and synthesis are still very much influenced by the linguists' phonetic analysis which determinedly inserts segment boundaries—an exercise that can be performed fairly consistently by those with suitable training, but an exercise that ignores the fact that many segmental boundaries are determined more by convention than acoustic reality. Those with a more phonological mindset are much more concerned with the interdependency of successive segments and the succession of unsynchronised acoustic features. Where do you insert the boundary between, say, a stop and a following vowel to separate the acoustic features of the vowel from those of the stop, when important cues for the stop are embodied in the course of the formant transitions to the vowel, which don't even begin and end at the same time, and, as [33] showed, formant tracks are not consistent between different contexts. Boundaries in glide, vowel and liquid sequences tend to be even more arbitrary.

When Shearme and Holmes [33] examined the vowels in continuous speech in the hope that clusters characterising them would emerge, they found there was a complete absence of such clustering. Thus speaker classification, to the extent that it needs to use the formant structure of vowels in the classification process, depends—at least to some extent—on speech recognition, so that the underlying phonetic structure can be used to recover useful vowel data. This puts both speech recognition and speaker classification into an interdependent relationship. You can help recognition by using information about the speaker, but you need information about the speaker to help with recognition.

It is possible that, in the cue-reduced environment of the telephone speech experiments referenced above [18], that this mutual support is sufficiently reduced that the listener's attention becomes focussed on the primary task of recognising what has been said, even though the ostensible task is to recognise who might be the speaker. It is also possible that, within the "in-group", listeners have learned to pay attention to some cues at the expense of others, and the cues that are used are simply absent from the band-limited telephone speech. Such questions need to be answered.

Fant [27] pointed out that the sound spectrograph was limited to an upper frequency of 3400 Hz whereas an upper limit of at least 8000 Hz was needed for unambiguous comparitive description of unvoiced continuants and stops—a limitation with early research on recognition and synthesis. Voice quality—important in speaker classification—may need an even higher upper bound on the frequencies considered in assessing voice quality, because higher formant and other spectral cues are likely to be important. The interest in telephone quality speech arises because the telephone is ubiquitous, and forensic cases may have nothing other than sample of telephone speech. Also, by increasing the difficulty of the task in psychophysical experiments, there is a greater chance of producing statistically measurable results. However, the down side is that, since we don't really know what we are looking for, we may eliminate, or at least attentuate, the relative significance of the very factors we should actually be exploring.

Foulkes et al. [34] note:

> "it has also been shown that children learning different languages display subtle differences in the phonetic forms they use to realise a phonological category. For example, American and Swedish children aged 2-6 differ in place and manner of /t/ production, in accordance with differences found in the speech of American and Swedish adults [35]. Similar differences were found for vowel duration among the same children (page 2)" and "The (t) variants therefore involve subtle and highly complex differences in the co-ordination of oral and laryngeal gestures. (page 14)"

A useful additional tool for investigating and understanding the nature and importance of such differences in the speech of different speakers would be an articulatory synthesiser, since experimenters could use artificial stimuli, systematically varying such subtle cues under controlled conditions, to determine their perceptual effect. This author believes it is time to consider perceptual experiments using artificial stimuli that can closely mimic real human speech with a full spectral range.

A good quality articulatory synthesiser that is easily but appropriately controlled, and which is inherently restricted to the potential acoustic output of a real human vocal tract, with supporting models to provide a foundation for manipulating speech production from the lowest sub-phonetic

articulatory level up to the prosodic level of rhythm and intonation would go a long way to providing the tool needed for such perceptual experiments. As Cooper et al. [25] pointed out in connection with earlier speech research, there are "many questions about the relation between acoustic stimulus and auditory perception which cannot be answered merely by an inspection of of spectrograms, no matter how numerous or varied these may be"

The Gnuspeech system

It has been the goal of building such an articulatory synthesizer and the necessary supporting models that has formed the subject of ongoing research by the present author and his colleagues [36]. The synthesis research has been ongoing for many years, first in the author's laboratory at the University of Calgary, then in 1990 it became the subject of a technology transfer exercise (to the now-defunct Trillium Sound Research—killed by the demise of NeXT Computer), and is now available to all under a General Public License as Gnuspeech—an ongoing GNU project [37]. Significant components of the complete, successful experimental system for articulatory synthesis that was developed on the NeXT have been ported to the Macintosh computer under OS/X and work is also under way to port it to GNU/Linux.

The complete system has been described elsewhere [36], [38], [39], and the source code is available for both the NeXT and the Macintosh (which is being modified to compile under GNUStep for GNU/Linux—though this port is not yet complete). Suffice it so say here that the approach builds on work carried out by Fant and Pauli [40] and by Carré [41]. By applying formant sensitivity analysis, and understanding the relationship of the resulting "Distinctive Regions" (Carré's term) to the articulatory possibilities inherent in the human vocal apparatus, the control problem for an articulatory synthesiser has been largely solved. In addition, the speed of modern computers allows the necessary complex computations for artificial speech based on the waveguide acoustic tube model to be carried out at a higher rate than is needed for real-time performance. Thus a tool is now available that allows experiments with the timing and form of articulatory events—with the caveat that the transformation between explicit articulatory specifications (such as tongue and jaw movements) and the Distinctive Region Model (DRM) equivalents has not yet been implemented, though the transformations are considered to be relatively straightforward.

Possibilities and current limitations of the experimental system

The articulatory synthesis tools that have been developed do enable significant experiments on the effects of learned speech articulations to be performed. The tools could be improved and extended in a number of ways to make such work easier by enabling easier control of some of the characteristics to be investigated (for example, by implementing the transformation between articulator movements and the DRM equivalents). The tools and interfaces were only an initial implementation of what was needed to develop databases for a complete English text-to-speech system based on the articulatory synthesiser, rather than full a psychophysical/linguistic laboratory tool-set. However, the system as it stands allows experiments with the timing of articulatory events to be performed, based on the observation that the DRM captures the essence of human articulatory possibilities. It has been used already to look at geriatric articulation and the timing of stops and stop bursts [42].

Perhaps most importantly, since the system provides a complete text-to-speech system based on better, effective models of speech production, rhythm and intonation, experiments on particular characteristics will be embedded in a context that provides natural variation of *all* formants, with good rhythm and intonation and with accurate records of what variations were used. Making all variations, rules and data involved in any synthesis formally explicit and editable was an important goal of the system development.

An important limitation of the system is the reality that it is actually still a hybrid system. Though the acoustic tubes representing the oral and nasal cavities give a true simulation of the acoustic behaviour of the appropriate human anatomy, with higher formants properly represented and variable, with

inherently correct energy balances, and with simulation of oral and nasal radiation characteristics, the larynx waveform is injected directly—albeit from a wavetable that can be dynamically varied—and the fricative, aspiration and other noises (such as bursts) are also injected at appropriate places in the tube model. This latter arrangement provides the basis for appropriate fricative formant transitional behaviour but the spectra of the injected noises are generic and approximate rather than individual and detailed. A better model would emulate the vibrating vocal folds, and oral tract constrictions, to generate the glottal waveform and noise spectra aerodynamically, based on accurate physiological models. The properties of all these noise spectra are characteristic of individual speakers.

The rhythm and intonation components of the system are based on the work of Jones [43], Pike [44], Jassem [45], Lehiste & Peterson [46], Abercrombie [47], Halliday [48], Allen [49], Ladefoged [50], Pierrehumbert [51], and Willems et al. [52], amongst many others as well as work in the author's laboratory at the U of Calgary. Wiktor Jassem spent a year in that lab and the results of the joint rhythm studies carried out are reported in [53] and [54]. Some of the intonation studies are reported in [55], [56] and [57]. Subsequent unpublished work on the intonation patterns has achieved significant improvement by using smoothed intonation contours, based on the timing events suggested by Allen [49]. Note that the original formant synthesizer used in our early research was replaced in 1993-4, by the articulatory synthesizer described in [36], previously cited.

Face recognition as an analogue of the speaker recognition problem

Zhao and his colleagues [58] have provided a good summary of the state of the art in facial recognition. They conclude that face recognition is a dedicated process that is separate from normal object recognition and that both wholistic and feature recognition strategies play a part, and facial expression seems to be recognised by separate mechanisms (somewhat as identity and location of objects are processed by different mechanisms in visual processing generally). Hair, face outline, eyes and mouth are significant, while the nose appears to play an insignificant role. Low spatial frequency components, bandpass components, and high frequency components seem to play different roles, with low frequency components permitting judgements concerning sex, but high frequency components being needed for individual identification. Other factors play a role, including lighting direction and being able to observe moving images rather than still photographs. Such observations may contain clues concerning how to approach speaker recognition and classification, with the major observation that the process is undoubtedly more complex than might be thought, and almost certainly involves different specialised mechanisms performing different tasks. Dynamic aspects are almost certainly important, and some kind of functional feature analysis, in addition to wholistic measures, is likely to help.

Simply taking statistical measures of energy variation in the spectrum, or pitch values, and the like, is akin to trying to recognise faces from photographs based on a statistical comparison of pixel characteristics (spatial frequencies, distribution of pixel densities, and the like), without trying to identify features such as hair, eyes, mouth and so on, as well as relevant dynamic clues. It is the dinosaur bone problem. If you don't take account of the underlying structure of the data, your statistics become too unfocussed to relate to reality in any precise way. As the Zhao et al. [58] state, quite clearly, for face recognition:

> "Feature extraction is the key to both face segmentation and recognition, as it is to any pattern classification task. For a comprehensive review of this subject see [59]."

It is also worth noting that the ability to observe a speaker's face affects the listener's ability to understand what the speaker is saying—from the fused perception of the McGurk effect [60] to the extreme form of lip-reading. Speech recognition is clearly multi-modal which, if nothing else, helps to illustrate some of the complexity of perception, and indicates that speaker classification is also likely multimodal to the extent that cues other than voice are available.

In achieving good mimicry of a speaker, whether by voice alone, or using additional cues such as facial expression and body language, the speaker mimic needs to do more than imitate voice quality, intonation, and accent. The mimic succeeds best if he or she captures the appropriate "persona" of the

person being mimicked—cool, excited, in control, sympathetic and so on—relating closely to *how the target would be expected to act* in the same circumstances.

Perception is a complex, active, organising process based on assumptions that work, in the real world. It is not passive. We see the moon as increasing in size, the nearer it is to the horizon because we are increasingly compelled to see it as increasingly far away. The fusion of the McGurk effect (hearing /b/, seeing /k/ and perceiving /d/, for example) arises because we have to reconcile the sound we hear with the conflicting appearance of the speakers lips and jaw. Close the eyes, and we perceive the sound that was actually produced. This is not the place to become diverted into a treatise on perception, but its active, organising nature is well documented in the literature. There is no reason to suppose that our approach to recognising speakers is any different, whether the categorisation is broad or narrow. The extent we succeed or fail in the task is a measure of the cues to which we learn to pay attention and those we learn to ignore, just as with learning to recognise the sounds of our native language as infants [61].

Back to the main goal—speaker classification

In his survey of speaker recognition, Atal ([14] p 460) Asks: "How do listeners differentiate among speakers?" and states that a satisfactory answer is not easy. In his review of automatic speaker verification in the same special issue of the IEEE proceedings, Rosenberg ([15] p 480) says that, for foils, mimicking behaviour and learned characteristics is less successful than obtaining a strong physiological correlation, but then quotes an experiment showing that even an identical twin was unable to imitate the enrolled sibling well enough to get accepted by a verification system when attempting to foil the system.

This tells us three things. First that, even in 1976, verification techniques were amazingly effective; secondly that possessing identical physiology did not give the advantage that might have been expected, given his earlier remarks; and thirdly, that the verification methods used must have captured some aspects of the speaker twins other than physiologically determined characteristics—somewhat refuting the notion that physiology was the core characteristic for discrimination. It also tells us that we have to look at learned speaking behaviour and other factors even for speaker recognition and verification, let alone for speaker classification.

Chollet and Homayounpour [7] carried out an extensive study to test the ability of listeners to discriminate the voices of twins. Family members were significantly better at the task than listeners not familiar with the twins, and the latter did not perform significantly differently from the two automatic procedures based on low-level acoustic features that were also tested. The authors conclude, amongst other things, that a speaker verification system which takes account of a speaker's behavioural characteristics will be more robust against foiling by a twin with a similar voice.

How does the problem change if, instead of wishing to verify an enrolled speaker, or identify a speaker from a group of speakers, the task is to determine something about the speaker such as age, sex, emotional state, the speaker's feeling of confidence, and so on? The reasons for wanting such information can be quite varied. Müller, in his thesis and recent papers, [62], [63] and [64] describes his AGENDER system, designed to obtain information about age and sex of speakers to allow an automatic shopping assistant to help a customer more effectively by tailoring its purchase recommendations based on the information extracted. A more ambitious goal is to understand the cues and behaviour in speech for purposes of synthesis, to create more realistic artificial agents. The German Research Center for Artificial Intelligence in Bremen has a project to create a "Virtual Human" [65]:

> "Creating a virtual figure as a conversation partner requires detailed, anthropomorphic design of the character, realistic speech, and emotional interactions, as well as, the exact simulation of movement in real time." (from the web site)

Such ambition goes beyond the scope of this chapter, but illustrates the directions of research of interest for both speaker classification and speech synthesis, and ties together the synthesis of speech, facial expression and body language—a topic that has also been of interest in this author's lab [66]

and [67]. It also parallels the work at the US Air Force Research Lab in Mesa, Arizona [2] previously cited.

Understanding the basis for adapting to speakers, according to their condition, type and situation, and responding appropriately, are increasingly important as machines become more involved in significant dialogues with people. The many reasons that people detest current voice reponse systems, comprise their one-size-fits-all approach to dialogue, coupled with painfully slow exploration of many possible choices, most of which are irrelevant to the caller, together with their total lack of natural dialogue and empathy, including their inability to assess urgency, puzzlement, or other dialogue conditions, as well as their stereotypical and mechanical approach to even the lowest levels of social nicety. Machines currently exhibit low emotional intelligence [68]—at least partly because they have little basis for performing appropriate speaker classification at present.

Although there has been considerable success in using multidimensional classification methods on "feature vectors" derived from various kinds of speech analysis, less work has been done on approaches to classification involving the explicit extraction of features known to be associated with the distinctions that the classifier is expected to make. In order to extend this work more information concerning such features and their relevance is needed. Given that—unlike identification, verification, sex or age determination—the judgements are less objective, it is necessary to understand the subjective aspects of speaker classification, at least as a precursor to or test for the development of more objective measures.

Intonation and rhythm:

Pitch has been successfully used in speaker identification and verification, but only at a statistical level, without a lot of attention to its *functional* patterning—where "functional" includes involuntary effects (such as the effects of fear on larynx performance) that would be important for more general classification tasks. It would be hard to get ethics approval for an experiment in which people were made so afraid that it affected their voice, and perhaps just as hard to make people that afraid in an experimental setting. This is an example where a good synthetic speech emulation of relevant factors could produce output for judgement by listeners as a means of exploring the markers for fear. Of course, this raises the question of what listeners are able to perceive versus what changes occur when the speaker is afraid. Given the reports already cited concerning speaker identification, it would seem that even trained listeners do not hear differences in speech characteristics that are measurable. At the same time, not all measurable differences are necessarily relevant to either speech recognition or speaker classification.

It is worth reiterating that picking up clues relevant to speaker classification may be made easier if the speech can be recognised, just as recognising speech may be made easier if there has been at least some degree of speaker classification. For example, recognising an accent is likely to be enhanced by a procedure that identifies vocalic segments, or recognises words that often contain glottal stops substituted for other stops in particular accents/dialects. Reductionism is no longer the best approach.

The larynx, which creates pitch pulses, functions at both segmental and suprasegmental levels, and both levels are relevant to various aspects of speaker classification. At the segmental level, for example, relative variation in voice onset time (VOT) in the transition from voiceless stops to voiced segments or for initial voiced stops can provide information relevant to linguistic background, sex and age [69]. Much of the work to date has focussed on the relevance of VOT at the segmental level, rather than as a cue for speaker classification. Such a measurement depends on identification of the segments concerned—that is, on speech recognition.

At the suprasegmental level, the frequency and amplitude of pitch pulses vary, and give information about intonation pattern and rhythm. These represent some of the dynamic features of speech in which we need to take an interest for speaker classification. By far the most important determinant of rhythm is the relative *duration* of the underlying segments—particularly nuclear vowels—which

are also associated with significant features of pitch change. The precise timing of the pitch changes relative to the segmental structure is almost certainly a useful clue to speakers and their characteristics (see also [49]). Again, speech recognition and speaker classification characteristics are mutually supportive and greater success in either is likely to be achieved if both are done in concert.

It is not clear to what extent the same is true of jitter (short-term pitch period variation) and shimmer (short-term pitch amplitude variation) nor how are these might be affected by rage, nervousness, age, illness, and so on. If they have relevance, is the effect more pronounced at semantically significant or phonetically significant places in the utterance. We simply don't know.

Modelling rhythm and intonation, even in general, have proved to be contentious topics for decades, and there is a plethora of different approaches to characterising intonation patterns and describing rhythm. One of the more obvious splits on rhythmic description is between those who consider English to have a tendency towards isochrony (equal durations between "beats" [47] and [48]), and those who say such a phenomenon is a fiction—an artifact of perception and the phonetic structure of words. Though I have not seen anyone say this explicitly, it could be that this is a difference between American English and British English. Certainly we found that "a tendency towards isochrony" accounted for 10% of the variance in segment duration in the body of British RP English that we examined in detail [53], [54], and our results have more recently been supported by work at the Centre for Speech Technology Research in Edinburgh [70]. The degree of "tendency towards isochrony" is another potential characteristic that may help categorise speakers. There are other aspects of rhythmic patterning and rate of speech that provide cues to the speaker class and condition such as pause patterns (think of Churchill, the wartime UK prime minister, and the way he spoke).

Intonation patterns provide a varied and shifting target since there seem to be many varieties of English intonation, and new forms arise before the old models have been evaluated and tested—a case in point being the arrival of "up talk" in which a pattern with rising pitch at the end of utterances seems to have become very popular, and supposedly derives from the "valley talk" which had its genesis in the 1960s in California. Other intonational patterns, if recognised and described, would provide valuable information concerning speaker class and condition. Again, think of Churchill's intonation as well as his rhythm when indulging in his famous oratory.

The question is not so much: "What is a good model of British or American English intonation?"; as "What resources are available to speakers, and how can these be clearly characterised in order to detect similar patterns in different groups of speakers, or determine their emotional and other psychological states?". What Eckman and Friesen [71], [72], did for the face we must do for speech, at least for intonation and rhythm, which seems the obvious place to start. Prosody (rhythm and intonation) probably plays an auditory role akin to that played by facial expression and body language in the visual domain. Humans have a peculiar sensitivity to facial expression—which is what prompted Levine [73] to use facial caricatures developed by Chernoff [74] to present multidimensional statistical data, as illustrated in Figure 1.

The choice of correspondence was done so insightfully that not only are the expressions produced very appropriate to the data being illustrated, but the editors felt compelled to disclaim any suggestion that the expressions indicated the mood of people in the cities from which the data were drawn. Prosody may offer similar possibilities in the auditory domain but we simply don't know. We do know that if analogous correspondences were found, we should have found a powerful possibility for characterising speaker state from prosody. Many readers will be familiar with the voice of Marvin "the metal man" in Douglas Adams' *Hitch-hikers Guide to the Galaxy* which provided an excellent auditory caricature of depression.

It would be interesting to perform experiments to characterise, and determine human sensitivity to, "tone of voice"—which would include both rhythm and intonation, as well as voice quality and facial expression. Understanding "tone of voice" in a formal sense would be an important step in dealing with a range of speaker classification tasks.

People with hearing impairments have identifiable differences in their rhythm and intonation—in fact, people with total hearing loss must undergo regular speech therapy to keep their voices in a reasonably normal state.

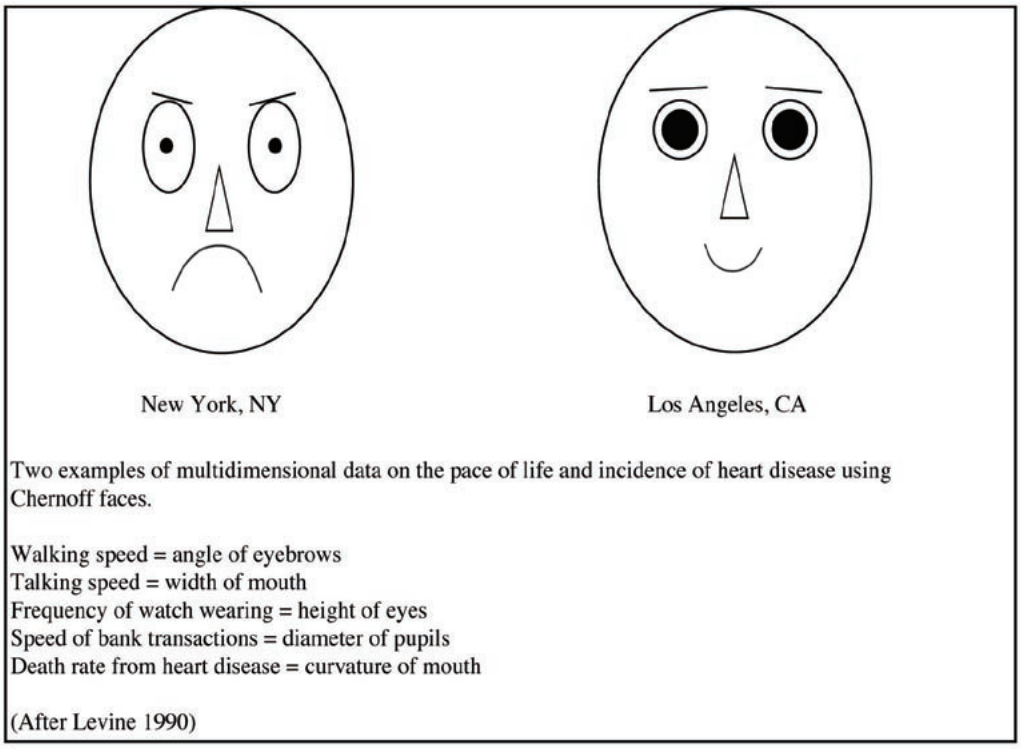

Figure 1: *Chernoff faces showing the pace of life and heart disease*

Abberton was one of the pioneers of speaker identification using solely information concerning intonation patterns [75]. She pointed out that intonation not only contains useful information for speaker identification, but also contains considerable information relevant to speaker classification. She included synthetic stimuli in the listening trials of her experiment to control for potential confounding factors. She quotes earlier experiments by Brown, Strong and Rencher [76] who also conducted listening experiments with synthetic speech to investigate the relationship between perceived "benevolence" and "competence" speaker characteristics. It is reasonable to suppose that clues may be obtained that relate to many factors such as: anger, enthusiasm, ethnicity/native language, fear, urgency, uncertainty, lying, submission, puzzlement, frustration, aggression, dominance and confidence. Both analytical and synthetic experiments would be appropriate. The ability to caricature any of these factors in synthetic speech would, as noted, be dramatic and informative. To the extent that subjective evaluation is important, hypotheses derived on the basis of analysis are best tested and validated by perceptual experiments using synthesis, provided the parameters of interest can be controlled in a reasonable way.

Lower level cues: segmental level and below

Suitable speaker verification/identification techniques undoubtedly extract measures closely related to the shapes and rates of formant transitions. These are characteristic of individual speakers who have learned speech habits. However, although this ability may serve its purpose, it does not contribute to knowing how to categorise speakers as opposed to how to recognise or verify them, partly because the nature of the features extracted are hidden inside the complexities of the automatic decision procedures that are the norm for such tasks these days; and partly because even if they were not hidden, or could be discovered, the information is not structured by, nor related to any knowledge of particular classification categories being sought. Determining that a particular speaker is who he

or she claims to be, or identifying which individual in a group is the one speaking is not the same as assigning such a speaker to any one of the large variety of possible categories listed at the end of the previous section (5.1), which is not exhaustive.

The question, as for higher level features, is not how do individuals differ in their acoustic characteristics when speaking, but in what ways are the members of some category of interest similar. This is a very different problem. To solve the problem requires that we examine the speech of in-group and out-group speakers, formulate hypotheses about similarities and differences, and then test these in various ways, including by synthesising speech with and without the characteristics that seem to identify in-group versus out-group individuals. Systematic psychophysical experiments can also be used directly as a way of finding ut what affects perception that the voice belongs to a particular group, just as perceptual tests in the early days allowed researchers to find out what acoustic characteristics were essential for the perception of particular categories of speech sound.

Some characteristics at the segmental level that may be of interest for categorising speakers include: relative formant amplitudes; rates and shapes of formant change; rates and shapes of articulatory movements (closely related to the previous item, but wider); formant ratios and values in known vowels; quality of vowels in known words (degree of reduction, actual formant values …); segment durations & statistics; segment ellipsis & substitution; use of markers such as glottal stops versus other stops; rate of speech; relative event timing at the segmental level (Voice Onset Time and stop durations are examples); spectral manifestations of sinuses and other physiological structures; nasalisation; and nasal spectrum. Again, note that speech recognition is an essential adjunct to extracting the features relevant to speaker classification.

In their paper on vowel clustering, already cited [33], Shearme and Holmes identified three generalisations concerning vowel formants, apart from the lack of signs of clustering. One was that plotting the formant tracks for a given speaker for each vowel produced could be used to draw relatively small areas containing at least a small portion of every track. The second was that these areas were significantly different for each speaker. The third was that each speaker's derived vowel-track F1-F2 areas were considerably displaced from the F1-F2 areas for the same vowels they produced in isolated monosyllables.

In a related but different experiment, using Lawrence's Parametric Artificial Talker—PAT [26], Broadbent & Ladefoged [77] synthesised sentences with different mean formant frequencies, all saying: "Please say what this word is". They also synthesised single-word stimuli "bit", "bet", and "bat". The single-word stimuli were presented to listeners, accompanied by different versions of the sentence, in an experimental design that provided an hour's intelligence-testing between two presentations of a test sentence and a stimulus word. There were seven different groups of subjects in which the details of the sentence/stimulus-word presentation varied—especially the delay between each sentence and the stimulus word. Most groups heard the sentence, followed—after a delay (depending on the group)—by the stimulus word. In one group, the stimulus word was presented first. The latter group showed little effect of the sentence on perception. A second group that counted during a 10 second delay also showed little effect.

Except for the conditions noted, it was found overall that the way speakers categorised the stimulus words as "bit", "bet" or "bat" was related in a simple way suggestive of a typical perceptual adaptation effect to variation in the mean formant frequencies of the preceding sentence. The same stimulus would be perceived as a different word by the same listener, depending on the formant frequencies of the preceding sentence. There were some unexplained anomalies

These experiments suggest: first, the actual spectral quality of the vowels is less important than the dynamics of the formant transitions from point of view of recognition; and secondly, if the appropriate small areas containing at least part of all a speaker's vowel formant tracks could be determined, these could be powerful clues to speaker classification—or at least speaker verification/identification.

Dynamics and longer term effects

Adami et al. [78] comment that: "Most current state-of-the-art automatic speaker recognition systems extract speaker-dependent features by looking at short-term spectral information. This approach ignores long-term information that can convey supra-segmental information, such as prosodics and speaking style." Their system, which uses Gaussian Mixture Modeling claims 3.7% error rates in speaker recognition (presented as a 77% relative improvement over other approaches), and they plan work on formants. This represents a small departure from the common obsession with "feature" selection, as opposed to looking at function and underlying mechanisms, even though their goal is only speaker recognition rather than classification.

Part of recognising a speaker is the dynamic *interactive* aspect—how they react in dialog, what choice of words and argument structure they use, how they signal how they are feeling, and so on. Similar characteristics are likely relevant to classifying speakers into groups but we need to understand how all these potential markers relate to the groups in which we are interested.

Recognising speakers gender and age, and sexual orientation

Carlson et al. [79] noted that "Special effort is invested in the creation of a female voice. Transformations by global rules of male parameters are not judged to be sufficient. Changes in definitions and rules are made according to data from a natural female voice." Such differences arise at both the segmental [68] and suprasegmental levels.

In producing convincing female speech from the *Gnuspeech* synthesiser, we found similar problems. Early in the development, Leonard Manzara produced three versions of the utterance: "Hello". By adding "breathiness" to the glottal excitation (one of the available utterance rate parameters) and by judiciously crafting the intonation and rhythm, a reasonably convincing female voice version was produced, using the standard rules for the segmental level synthesis. The male and child voices were less trouble, though the child voice is probably more like a boy than a girl. All the voices could probably be improved if we understood the markers better, and this would provide a better basis for making these important categorisations is speaker classification. The relevant speech synthesis examples are provided for listening in connection with [36] which is available on-line. A great deal more understanding of the differences between male, female, and child voices, as well as more general markers for age, is required—research that is likely best carried out using an articulatory synthesiser similar to the *Gnuspeech* system in concert with careful study of relevant spectrographic data and previous research.

It seems probable that the markers involved in these kinds of distinction also play a part in the voice quality and intonation often associated with speakers with specific sexual orientation—for example, some gay men. By casting the research net wider to encompass such speaker categorisation, even more should be learned about the resources *all* speakers use to project their identity through speech, and assist with speaker classification.

Conclusions

A major conclusion is that speaker classification requires the isolation of features relevant to specific kinds of categorisation task, and that many of these features can only be extracted on the basis of a reasonable capability for recognising what has been said—that is, by speech recognition—and by using other knowledge about the structure of speech, with better ways of characterising the resources used for such speech attributes as rhythm and intonation. Without such informed structuring of the data, and identification of the linguistic and paralinguistic structure, any statistics that are derived may allow reasonable success at identifying or verifying particular speakers (on the same principle that photographic comparisons may allow people to be identified or verified on the basis of pixel images), but the "bones" that have been identified will be very hard to classify into meaningful groups of the different kinds needed for useful speaker classification.

A second conclusion is that, for any speech or speaker recognition task, much greater *explicit*

attention should be paid to *dynamic* properties of the speech signal, at both the segmental and the suprasegmental level. Additionally, dialogue structure information—also dynamic—can provide important information.

A third major conclusion is that only by paying attention to the underlying structure of speech, explicitly, shall we continue to make progress in both speech recognition and speaker classification. Ignoring the dinosaurs whose bones we are examining will only take us so far. Most modern pattern classification approaches deliberately hide this underlying structure inside automatic methods, if it is used at all. We need to expand our approaches and stop focussing on reductionist solutions.

Hollien [80] opens the section on speaker identification in his book by saying:

> "Almost anyone who has normal hearing, and who has lived long enough to read these words, has had the experience of recognizing some unseen speaker (usually someone familiar) solely from listening to his or her voice. It was from this everyday experience that the concept (or is it a myth?) of speaker identification was born."

Very similar remarks could be made about the everyday experience of judging mood, ethnicity, intent, and a host of other factors relevant to speaker classification just from hearing someone speak—whether seen or unseen. Campbell [5, p 1446] refers to "the curse of dimensionality" referring to the problem that automatic feature extraction (as in the speaker identification task) soon causes resources to be overwhelmed unless some kind of statistical model is used to manage and structure the plethora of data. This paper points out some of the avenues and directions towards which research for relevant structure in speaker classification may usefully be directed, and reminds the reader of the importance of experiments with synthetic speech in this quest.

Acknowledgements

The author would like to thank the Natural Sciences and Engineering Research Council of Canada for providing financial support for his research until 1992 under Grant OGP0005261. He would also like to thank the shareholders and principals of Trillium Sound Research Inc. for agreeing to release all claim on the *Gnuspeech* software and other materials related to the work so that they could be made available to the speech research and lingusitics communities under a Free Software Foundation "General Public Licence" (GPL) (see: http://www.gnu.org/copyleft/gpl.html).

References

1. Arkin, W.M.: When seeing and hearing isn't believing. Washington Post, February 1 (1999), http://www.washingtonpost.com/wp-srv/national/dotmil/arkin020199.htm
2. Shockey, L., Docherty, G., Foulkes, P., Lim, L.: foNETiks (A network newsletter for the International Phonetic Association and for the Phonetic Sciences (August) "Research Scientist Software Engineering Air Force Research Laboratory, Lockheed Martin Operations Support, Mesa, AZ" (2006)
3. Foulkes, P.: The social life of phonetics and phonology. University of York, Aberdeen Symposium, 12 September, Power Point Presentation: http://www-users.york.ac.uk/~pf11/aberdeen05-webversion.ppt (136 slides) (2005)
4. Bimbot, F., Bonastre, J-F., Fredouille, C., Gravier, G., Magrin-Chagnolleau, I., Meignier, S., Merlin, T., Petrovska-Delacrétaz, Reynolds, D.A.: A Tutorial on Text-Independent Speaker Verification. EURASIP Journal on Applied Signal Processing **4** (2004) 430–451 (http://www.hindawi.com/GetArticle.aspx?doi=10.1155/S1110865704310024)
5. Campbell, J.P. Jr.: Speaker recognition: a tutorial Proceedings of the IEEE **85** (9), Sept (1997) 1437-1462
6. Chollet, G., Homayounpour, M.: Neural net approaches to speaker verification: comparison with second-order statistical measures. Proc. ICASSP 95 (1995)
7. Homayounpour, M., Chollet, G.: Discrimination of the voices of twins and siblings for speaker verification. Proc. EUROSPEECH '95, 4th. European Conference on Speech Communication and Technology, Madrid, Sept. (1995) 345-348

8. Genoud, D., Chollet, G.: Segmental approaches to automatic speaker verification. Digital Signal Processing: a Review Journal (2000)
9. Huang, R., Hansen, J.H.L.: Unsupervised Audio Segmentation and Classification for Robust Spoken Document Retrieval. Proc ICASSP 2004, Montreal, May 17-21, **1** (2004) 741-744
10. Lapidot, I., Guterman, H.: Resolution Limitation in Speakers Clustering and Segmentation Problems. 2001: A Speaker Odyssey, The Speaker Recognition Workshop, Crete, Greece, June 18-22 (2001) 169-173
11. Meignier, S., Bonastre, J-F., Magrin-Chagnolleau, I.: Speaker utterances tying among speaker segmented audio documents using hierarchical classification: towards speaker indexing of audio databases. Proc. ICSLP 2002 (2002)
12. Meinedo, H., Neto, J,: A stream-based audio segmentation, classification and clustering pre-processing system for broadcast news using ANN models. Proc. INTERSPEECH 2005 (2005)
13. Sanchez-Soto, E., Sigelle, M., Chollet, G.: Graphical Models for Text-Independent Speaker Verification. *in* "Non Linear Speech Processing" (Chollet G, Esposito A, Faundez M and Marinaro M, eds.) Springer Verlag, LNCS **3445** (2005)
14. Atal, B.S.: Automatic recognition of speakers from their voices. Proc. IEEE **64** (4), (1976), 460-475
15. Rosenberg, A.E.: Automatic speaker verification: a review. Proc IEEE **64** (4) (1976) 475-487, April
16. Sandmel, S. (general editor): The New English Bible with apocrypha (Oxford Study Edition) Oxford University Press: New York, Judges Chapter 12, verses 5-6 (1976) 266
17. Shirt, M.: An Auditory Speaker-Recognition Experiment Comparing the Performance of Trained Phoneticians and Phonetically Naive Listeners. Leeds [England] Working Papers in Linguistics. **1** (1983) 115-7.
18. Foulkes, P., Barron, A.: Telephone speaker recognition amongst members of a close social network. Forensic Linguistics **7** (2) (2000) 180-198
19. M4 Project: Annual report. (2004) http://www.m4project.org/M4-AnnualReport2004/
20. McCowan, I., Gatica-Perez, D., Bengio, S., Moore, D., Bourlard, H.: Towards Computer Understanding of Human Interactions. *in* Euro. Symp. Ambient Intelligence (EUSAI), Eindhoven, The Netherlands (2003)
21. Ginzburg, J.: Dynamics and the semantics of dialogue. *in* Seligman, J. Westerståhl, D. (eds.), Logic, Language, and Computation, CSLI: Stanford, CA (1996) 221–237
22. McClelland, E.: Familial similarity in voices. Paper presented at the BAAP Colloquium, University of Glasgow (2000) April.
23. Shirt, M.: An auditory speaker recognition experiment. Proceedings of the Institute of Acoustics Conference, **6** (1984) 101-4
24. Koenig, W., Dunn, H.K., Lacy, L.Y.: The sound spectrograph. J. Acoust Soc Amer 18 (1946) 19-49, July
25. Cooper, F.S., Liberman, A.M., Borst, J.M.: The interconversion of audible and visible patterns as a basis for research in the perception of speech. Proc. Natl. Acad. Sci. **37** (1951) 318-325
26. Lawrence, W.: The synthesis of speech from signals which have a low information rate. in "Communication Theory" (Jackson W, ed.), Butterworth & Co: London (1953) 460-469
27. Fant, C.G.M.: Modern instruments and methods for acoustic studies of speech. Proc. 8th. Int. Cong. Linguists, Oslo U. Press: Oslo (1958)
28. Liberman, A., Ingemann, F., Lisker, L., Delattre, P., Cooper, F.S.: Minimal Rules for Synthesizing Speech. J. Acoust. Soc. Amer. **31** (1959) 1490-1499
29. Holmes, J.N.: Speech synthesis by rule. Language and Speech **7** (3) (1964) 127-143 July-Sept
30. Mattingly, I.G.: Synthesis by rule of prosodic features. Language and Speech **9** (1966) 1-13

31. Fant, C.G.M.: Acoustic Theory of Speech Production. Mouton: The Hague, Netherlands (1960)
32. Holmes, J.N.: Research on speech synthesis carried out during a visit to the Royal Institute of Technology, Stockholm Nov, 1960 - March 1961. [UK] GPO Eng. Dept. Report No. 20739 (1961) October
33. Shearme, J.N., Holmes, J.N.: An experimental study of the classification of sounds in continuous speech according to their distribution in the forman 1 - formant 2 plane. Proc 4th Int Cong Phonetic Sci, Helsinki 1961, Mouton: The Hague, Netherlands (1962)
34. Foulkes, P., Docherty, G.J., Watt, D.J.L.: The emergence of structured variation. University of Pennsylvania Working Papers in Linguistics **7** (3) (2001) 67-84
35. Stoel-Gammon, C., Williams K., Buder, E.: Cross-language differences in phonological acquisition: Swedish and American /t/. Phonetica **51** (1994) 146-158
36. Hill, D.R., Manzara, L., Taube-Schock, C-R.: Real-time articulatory speech-synthesis-by-rules. Proc. AVIOS 95 (1995)(14th. Ann. Int. Voice Technologies App. Conf. Amer. Voice I/O Soc., San Jose, Sept 11-14, AVIOS: San Jose (1995) 22-44 (http://www/pages.cpsc.ucalgary.ca/~hill/papers/avios95—includes samples of synthetic speech)
37. Hill, D.R.: (2003) GNUSpeech: Articulatory Speech Synthesis. A GNU project. (http://savannah.gnu.org/projects/gnuspeech/)
38. Hill, D.R.: Manual for the "Monet" speech synthesis engine and parameter editor (http://pages.cpsc.ucalgary.ca/~hill/papers/monman/index.html) (2002)
39. Hill, D.R.: Synthesizer Tube Resonance Model user manual (http://pages.cpsc.ucalgary.ca/~hill/papers/synthesizer/index.html) (2004)
40. Fant, C.G.M., Pauli, S.: Spatial characteristics of vocal tract resonance models. Proc. Stockholm Speech Communication Seminar, KTH: Stockholm, Sweden (1974)
41. Carré, R.: Distinctive regions in acoustic tubes. Speech production modelling. J. d'Acoustique **5** (1992) 141-159
42. Slawinski, E.: Various experiments carried out in the Psychology lab. at the U. of Calgary (unpublished)
43. Jones, D.: An outline of English phonetics. Heffer: Cambridge, UK (1918) 237-242)
44. Pike, K.L.: Intonation in American English. U. Michigan Press: Ann Arbor (1945) (reprinted 1970)
45. Jassem, W.: Stress in modern English. Bulletin de la Société Linguistique Polonaise **XII** (1952) 189-194
46. Lehiste, I., Peterson, G.E.: Some basic considerations in the analysis of intonation. J Acoust Soc Amer **33** (4), (1961)April
47. Abercrombie, D.: Elements of general phonetics. Edinburgh U. Press: Edinburgh, UK (1967)
48. Halliday, M.A.K.: A course in spoken English: intonation. Oxford U Press: London (1970)
49. Allen, G.D.: The location of rhythmic stress beats in English: an experimental study I, and II Language and Speech **15** (1972) 72-100 and 179-195
50. Ladefoged, P.: A course in phonetics. Harcourt Brace Jovanovich: New York (1975)(pp 102-103)
51. Pierrehumbert, J.: Synthesizing intonation. J Acoust Soc Amer **70** (4) (1981) Oct
52. Willems, N.J., Collier, R., 't Hart, J.: A synthesis scheme for British English intonation. J. Acoust. Soc. Amer. **84** (4) (1988) 1250-1261, Oct
53. Hill, D.R., Witten, I.H., Jassem, W.: Some results from a preliminary study of British English speech rhythm. Research Report Number 78/26/5, Dept. of Comp. Sci., U. Calgary, Alberta, Canada T2N 1N4, January (1978) (Presented in Session V. Speech Communication V: Linguistics and Prosodic Features at the 94th. Meeting of the Acoustical Society of America, December 1977)

54. Jassem, W., Hill, D.R., Witten, I.H.: Isochrony in English speech: its statistical validity and linguistic relevance. Pattern, Process and Function in Discourse Phonology (collection ed. Davydd Gibbon), Berlin: de Gruyter (1984) 203-225
55. Hill, D.R., Reid, N.A.: An experiment on the perception of intonational features. Int. J. Man-Machine Studies **9** (2) (1977) 337-347
56. Hill, D.R., Schock, C-R.: Unrestricted text-to-speech revisited: rhythm and intonation. Proc. ICSLP 92, Banff, Alberta Oct 12-16 (1992) 1219-1222
57. Taube-Schock, C-R.: Synthesizing Intonation for Computer Speech Output. M.Sc. Thesis, (awarded Governor General's Gold Medal), Dept of Computer Science, U. of Calgary (1993)
58. Zhao, W., Chellappa, R., Rosenfeld, A., Phillips, P.J.: Face Recognition: A Literature Survey, ACM Computing Surveys (2003) 399-458
59. Chellappa, R., Wilson, C.L., Sirohey, S.: Human and machine recognition of faces, a survey. Proc IEEE **83** (1995) 705-740
60. McGurk, H., MacDonald, J.: Hearing lips and seeing voices, Nature, Vol 264 (5588) (1976) 746-748
61. Kuhl, P.K.: Infants' perception and representation of speech: development of a new theory. Proc. ICSLP 92, Banff, Alberta, Oct 12-16 (1992) 449-456
62. Müller, C.: Zweistufige kontextsensitive Sprecherklassifikation am Beispiel von Alter und Geschlecht. Dissertation zur Erlangung des Grades Doktor der Ingenieurwissenschaften (Dr.-Ing.) der Naturwissenschaftlich-Technischen Fakultät I der Universität des Saarlandes (2005)
63. Müller, C.: Automatic Recognition of Speakers' Age and Gender on the Basis of Empirical Studies. INTERSPEECH 06 (2006)
64. Müller, C.: Classification Post-Processing Using Dynamic Bayesian Networks. Amer. Ass. Artificial Intelligence Conf. (2006)
65. DFKI: Anthropomorphic Interaction Agents. Deutsches Forschungszentrum für Künstliche Intelligenz GmbH, Federal Ministry of Education and Research, Germany (2006) (website) http://www.virtual-human.org
66. Wyvill, B.L.M., Hill, D.R.: Expression control using synthetic speech. Tutorial #26: The State of the Art in Facial Animation, SIGGRAPH 90, Dallas, Aug 6th-10th. (1990) 186-212 (with demonstration video: "Testing 1, 2, 3, 4")
67. Hill, D.R., Pearce, A., Wyvill, B.L.M.: Animating speech: an automated approach using speech synthesised by rules. The Visual Computer **3** (5) (1988) 277-289, Mar.
68. Goleman, D.: Emotional Intelligence: why it can matter more than IQ. Bantam Books, New York, Toronto, London, Sydney, Auckland (1995)
69. Stolten, K., Engstrand, O.: Effects of sex and age in the Arjeplog dialect: a listening test and measurements of preaspiration and VOT. Speech Technology Laboratory, KTH, Stockholm TMH-QPSR **44** (2002). Paper presented at Fonetik 2002, May 29-31, Stockholm, 29-32
70. Williams, B., Hiller, S.M.: The question of randomness in English foot timing: a control experiment. Journal of Phonetics **22** (1994) 423-439
71. Eckman, P., Friesen, W.: Unmasking the human face. Consulting Psychologist Press: Palo Alto, California (1975)
72. Eckman, P., Friesen, W. Manual for the facial action coding system. Consulting Psychologist Press: Palo Alto, California (1977)
73. Levine, R.V.: The pace of life. Amer Scientist **78** (5) (1990) 450-459
74. Chernoff, H.: The use of faces to represent points in a k-dimensional space graphically. J Amer Statistical Assoc **68** (1973) 361-368
75. Abberton, E., Fourcin, A.J.: Intonation and speaker identification. Language and Speech **21** (1978) 305-318.

76. Brown, B. L., Strong,W. J., Rencher, A. C.: Fifty-four voices from two: the effects of simultaneous manipulations of rate, mean fundamental frequency, and variance of fundamental frequency on ratings of personality from speech. J. Acoust. Soc. of Amer. **55** (1974) 313–318
77. Broadbent, D. Ladefoged, P. Vowel judgements and adaptation level. Proc Royal Soc Series B (Bio. Sci.) **151** (944) Feb 2nd (1960) 384-399
78. Adami, A.G., Mihaescu, R., Reynolds, D.A., Godfrey, J.J.: Modeling prosodic dynamics for speaker recognition. Proc. ICASSP 03 (2003)
79. Carlson, R., Granstrom, B. Karlsson, I.: Experiments with voice modelling in speech synthesis. Speech Technology Laboratory, KTH, Stockholm STL-QPSR 2-3 (1990) (Paper also presented as invited tutorial paper to the ESCA workshop on Speaker Characterization in Speech Technology, Edinburgh, June 26-28 1990)
80. Hollien, H.: The Acoustics of Crime: The new science of forensic phonetics. New York/ London: Plenum Press (1990)