A WHISTLE-STOP TOUR OF AUTOMATIC SPEECH RECOGNITION

Prof Thomas Niesler

12 November 2013

A whistle stop tour of automatic speech recognition

Inaugural lecture delivered on 12 November 2013

Prof Thomas Niesler
Department of Electronic Engineering
Faculty of Engineering
Stellenbosch University

Editor: SU Language Centre
Printing: SUN MeDIA
ISBN: 978-0-7972-1462-0

ABOUT THE AUTHOR

Thomas Niesler's academic career began at the University of Stellenbosch where he obtained the BEng and MEng degrees in Electronic Engineering in 1991 and 1993 respectively. He moved to St John's College, Cambridge, in 1994 as a Benefactor's Scholar and obtained his PhD from the University of Cambridge in 1998 on the subject of statistical language modelling for large vocabulary speech recognition. He was employed first as a research associate (1997) and subsequently as a lecturer (1998–2000) by the University of Cambridge, where he taught undergraduate and master's-level courses as part of the MPhil in Computer Speech and Language Processing. He joined the Department of Electrical and Electronic Engineering at the University of Stellenbosch in November 2000, where he teaches undergraduate and postgraduate students in signal and speech processing. In 2002 he was an invited researcher at the NTT Speech Open Laboratory in Kyoto, Japan. He has authored 22 journal publications, as well as 24 refereed international and a further 19 national conference publications. He holds a C2 rating with the South African National Research Foundation, and has supervised 16 postgraduate students to graduation. His research interests lie in signal and pattern recognition, with a particular emphasis on human language technology. He is a member of the International Speech Communication Association.

A WHISTLE-STOP TOUR OF AUTOMATIC SPEECH RECOGNITION

Dave: Hello, HAL. Do you read me, HAL?

HAL: Affirmative, Dave. I read you.

Dave: Open the pod bay doors, HAL.

HAL: I'm sorry, Dave. I'm afraid I can't do that.

Dave: What is the problem?

HAL: I think you know what the problem is just as well as I do.

Dave: What are you talking about, HAL?

HAL: This mission is too important for me to allow you to jeopardise it.

Dave: I don't know what you're talking about, HAL.

HAL: I know that you and Frank were planning to disconnect me, and I'm afraid that's something I cannot allow to happen.

Dave: Where the hell did you get that idea, HAL?

HAL: Dave, although you took very thorough precautions in the pod against my hearing you, I could see your lips move.

Dave: Alright, HAL. I'll go in through the emergency airlock.

HAL: Without your space helmet, Dave? You're going to find that rather difficult.

Dave: HAL, I won't argue with you anymore! Open the doors!

HAL: Dave, this conversation can serve no purpose anymore. Goodbye.

From *2001: A Space Odyssey*, produced and directed by Stanley Kubrick and Arthur C. Clarke, 1968.

Not long ago, man did not converse with machines. That has begun to change. In the following I will sketch the development of automatic speech recognition (ASR) over the last six decades, and point out current challenges.

1 SPEECH AS A SIGNAL

Human speech can be considered to be a pattern of frequencies in time. It is often visualised as a spectrogram, which is a graphical depiction of a signal in which the vertical and horizontal axes denote frequency and time respectively. The third dimension, denoting energy, is represented by colour. Figure 1 shows an example of a spectrogram, and indicates the location of the words as well as the phonemes that compose the utterance in question.

This example illustrates some of the challenges of automatic speech recognition. Firstly, there are in general no clear boundaries between either the words or the sounds. Instead, the sounds usually flow into one another smoothly. Secondly, consider the two instances of the word 'zero'. While one can see strong similarities between the two repetitions of this word, they are clearly not the same. Differences are apparent despite an ideal scenario: the words occur within the same utterance and are therefore produced by the same speaker in the same style and in the same recording environment.

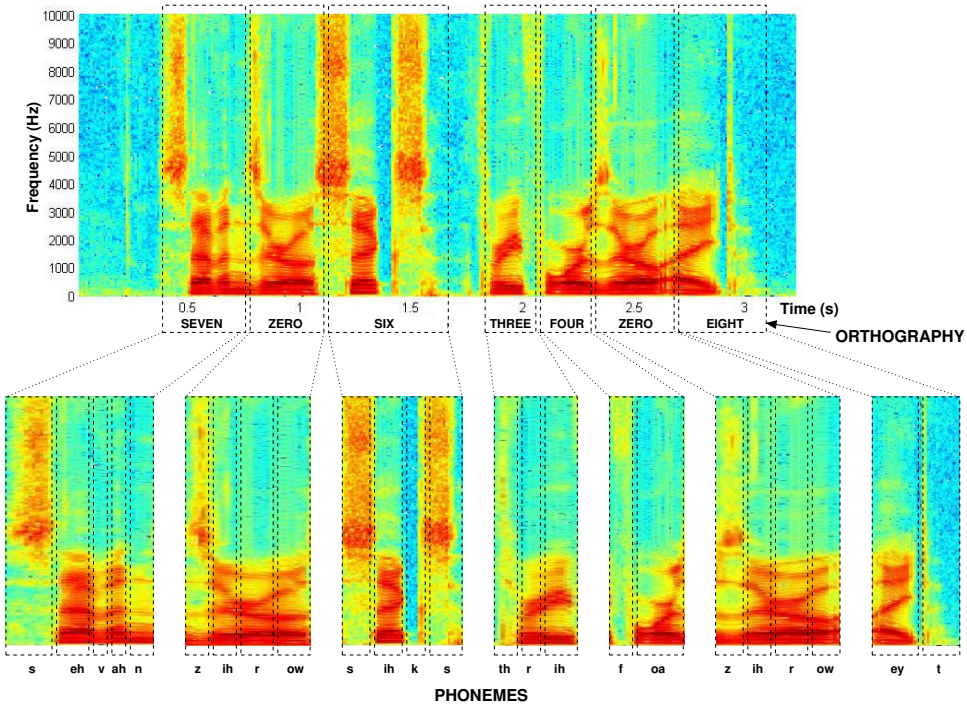

Figure 1: A spectrogram of the utterance 'seven zero six three four zero eight' and its decomposition into words and phonemes.

2 FIRST SYSTEMS: 1950–1970

Speech recognition technology was in its infancy in the 1950s. Among the very first systems was Audrey (Automatic Digit Recognizer), which was developed at Bell Laboratories in the USA in 1952 (Davis, Biddulph & Balashek, 1952). This system predates the advent of digital computers, and was completely built using analog electronic circuits. It understood the numbers 0 to 9 and required the speaker to pause for 350 milliseconds between words. A decade later, IBM showcased a similar device named the Shoebox, depicted in Figure 2. In addition to the ten digits, this machine could recognise six arithmetical command words.

Both Audrey and the Shoebox were highly speaker dependent, with accuracies dropping vastly when presented with a new voice. Then, in 1968, the world was introduced to HAL in Stanley Kubrick and Arthur C. Clarke's science fiction classic *2001: A Space Odyssey*. In stark contrast to Audrey and the Shoebox, in this film astronauts could converse fluently and naturally with a shipboard computer named HAL-9000.

Figure 2: The IBM Shoebox speech recogniser (Davis et al., 1952)

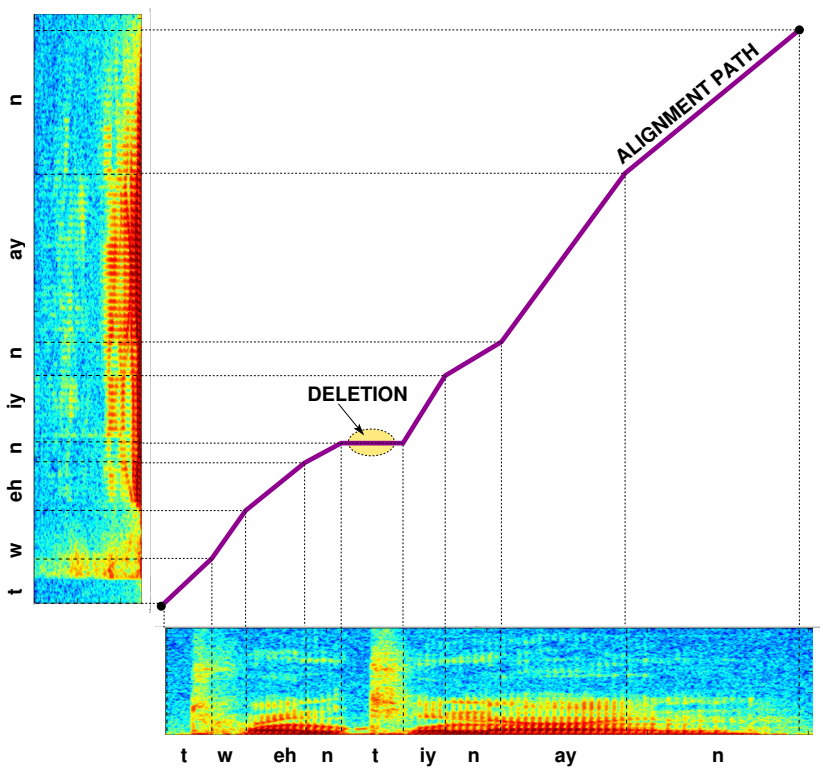

Figure 3: Alignment of two utterances of 'twenty nine' by dynamic time warping (DTW). Note the deletion of the second stop in the vertically displayed utterance.

3 DYNAMIC TIME WARPING: 1970–1990

The advent of digital computers led to the application of the dynamic programming principle (Bellman & Dreyfus, 1966) to the speech recognition problem. First proposed by the Russian scientist Vintsyuk in 1968, this yielded a technique now commonly referred to as dynamic time warping (DTW) or template matching (Vintsyuk, 1968). DTW was subsequently applied more extensively to ASR by Sakoe and Chiba in Japan in 1971 (Sakoe & Chiba, 1971; 1978).

When presented with two utterances of the same word, DTW accounts for the inevitable mismatch in the lengths and character of the constituent sounds by recursively computing the best alignment in time. This may be imagined as a process in which the time axes of the utterances are either 'stretched' or 'squeezed' in order for corresponding sounds to line up, as illustrated in Figure 3.

As a by-product of this alignment, DTW produces an 'alignment score', which is a numerical indication of how close the two waveforms are after the squeezing and stretching.

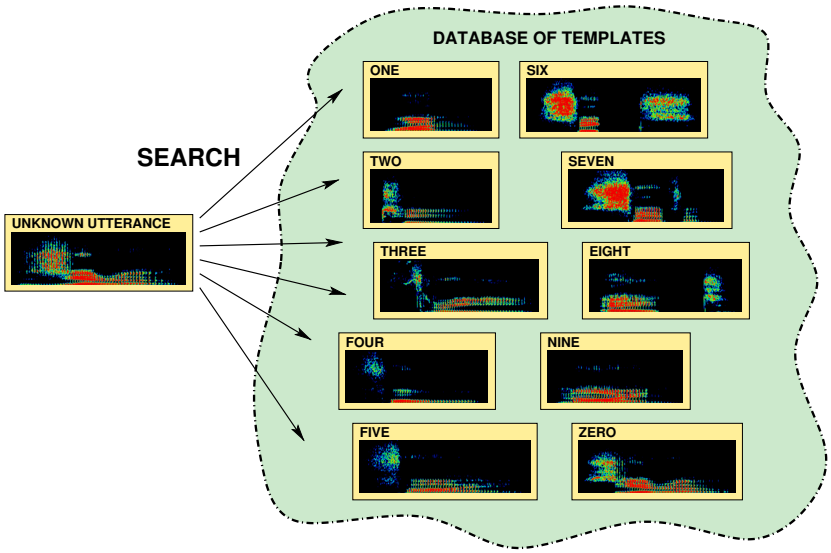

Figure 4: Isolated digit recognition by dynamic time warping (DTW).

This score can be used as an indicator of the closeness between a new user utterance, and each member of a database of template utterances with known transcriptions. By matching new speech against a set of stored templates in this way, speech recognition can be performed (see Figure 4). The application of DTW in ASR was quite successful, and allowed the field to gain momentum in the 1970s. However, it is naturally suited to the recognition of isolated words or phrases, and not to the recognition of unrestricted connected speech in which there are no pauses between words. Moreover, the computational load of the recognition procedure scales linearly with the number of templates. Today's speech recognisers are developed using hundreds of thousands of words, rendering the DTW approach completely infeasible.

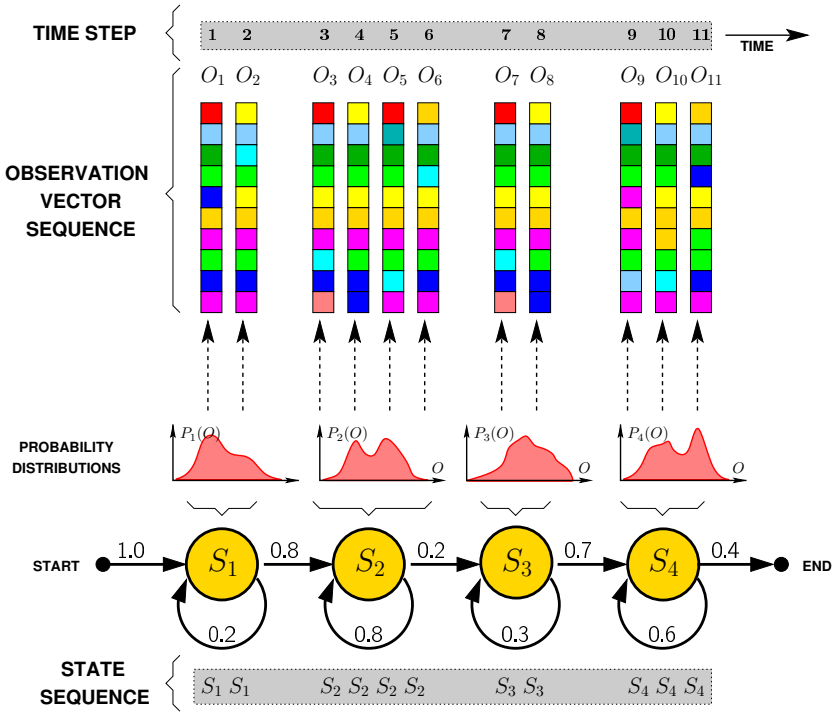

Figure 5: Illustration of a hidden Markov model (HMM) generating a sequence of observation vectors.

4 HIDDEN MARKOV MODELS: 1990 TO TODAY

Hidden Markov models (HMMs) are a generalisation of Markov chains, which were developed by the Russian mathematician Andrei Markov in the early 20th century (Markov, 1971). The extension of the Markov chain, with its directly observable states, to the hidden Markov model, in which the states are observable only indirectly through state-dependent probability distributions, was subsequently pioneered by Russian physicist Leontevich Stratonovich (1960) and later extended by American mathematician Leonard Baum and colleagues (Baum et al., 1970). Instead of storing multiple templates for the same word or sound, as would be required by a DTW-based approach, the HMM allows all templates for the same baseform to be absorbed into a single statistical model. This provides the enormous advantage of allowing a speech-recognition system to learn from virtually unlimited amounts data without increasing the computational cost at recognition time. However, it was not until the 1980s, when a classic tutorial paper was published by Larry Rabiner (1989), that the HMM approach became mainstream.

A hidden Markov model is a finite-state machine characterised by probabilistically weighted transitions between states and output observations emitted upon entering each state. It is particularly well suited to modelling sequential processes, such as discrete time series, and speech in particular. Figure 5 illustrates a four-state HMM. Each of the states S_i (i = 1, 2, 3 or 4) in the figure has an associated 10-dimensional observation probability distribution $P_i(O_j)$, which governs the generation of the observations O_j depicted at the top of the figure. The dimensionality of 10 for observation vectors is for the purposes of illustration only; a more typical value for speech recognition would be 39. A path would enter the first state, S_1, at time t = 1 and immediately lead to the generation of the observation vector O_1. At the next time instant, t = 2, the path must move to state S_2 with probability 0.8, or loop back to state S_1 with probability 0.2. For the example in Figure 5, it remains in S_1, generating observation vector O_2, and moves to S_2 at time t = 3, where it generates O_3. In this way the path propagates though the HMM, until it reaches the end by exiting state S_4.

What has just been described is the **generative** application of a HMM. For ASR, however, it is the **analytical** application of HMMs that is of interest. Assume that the observations O_j each represent a 'snapshot' of the speech signal at regularly spaced instances in time. Generally it is arranged that the observations capture the instantaneous spectral characteristics of the speech signal. Now consider any one of the observations $O_1 \ldots O_{11}$ in Figure 5. Because a state-dependent probability distribution $P_i(O)$ relates the observation O_j to the state S_i by which it was generated, it is not in general possible to unambiguously identify the state S_i given the observation O_j. As a consequence, for a sequence of consecutive observations, the generating state sequence can only be described in probabilistic terms, and is therefore *hidden*. Herein lies the usefulness of the HMM: a single model can account for an infinite number of different observation sequences by describing them probabilistically in terms of the model's transition probabilities and the state-dependent observation densities. By considering each possible path through the HMM, and the resulting probability that an observation vector sequence was generated by that particular path, one can calculate the probability that any particular observation vector sequence was generated by a HMM. These probabilities can be calculated efficiently using the Viterbi algorithm (Viterbi, 1967), and lie at the core of today's speech recognition systems. Each sound in a language is now modelled by a single HMM. Since finite-state models can be trivially interconnected, a word can be modelled by concatenating the HMMs of its constituent sounds using a linguistic resource known as the pronunciation dictionary, as illustrated in Figure 6. Furthermore, sequences of words can be interconnected according to a grammar of the language, leading to a final and invariably extremely large hidden Markov model. A Viterbi search through this network, given a sequence of observation vectors obtained from a new utterance, can find the path through the network which is most likely to lead to the observations. By tracing back along this path, the sequence of words, phonemes and finally states of which it consists can be determined. This sequence of words is the final result of the ASR process: the recognition hypothesis.

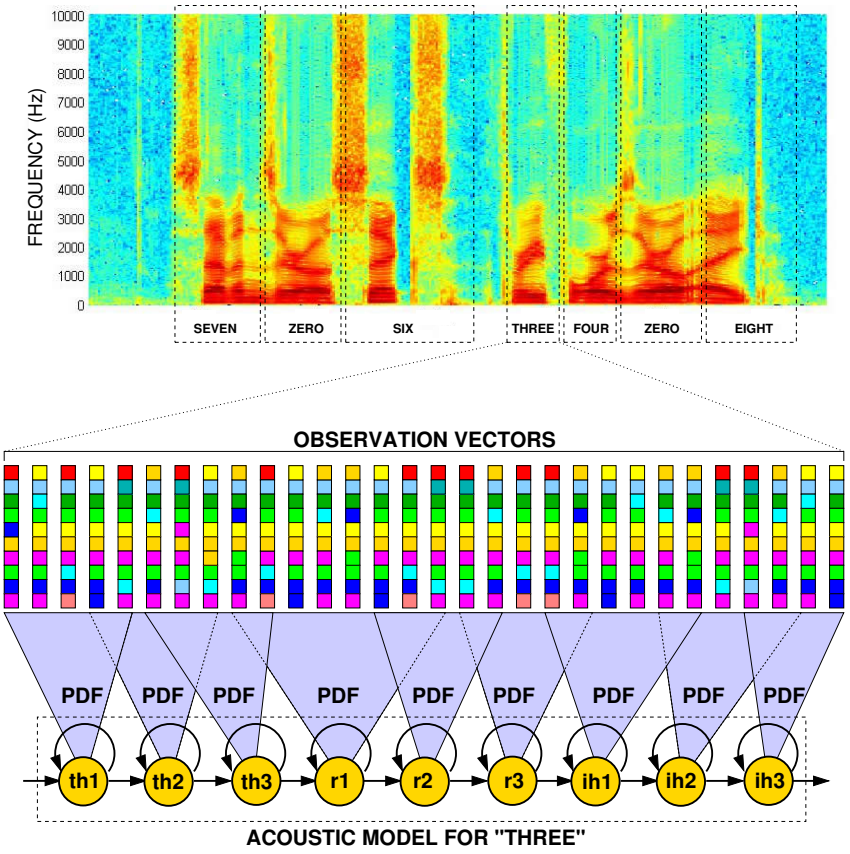

Figure 6: Whole-word acoustic modelling by concatenation of phoneme hidden Markov models (HMMs). Three HMM states are used to model each of the constituent phonemes.

5 THE FRONTIERS: TODAY AND TOMORROW

The strength of the hidden Markov model for ASR is its ability to absorb variability into a single model. This is necessary because the signal associated with a particular word or sound is highly variable, and this in turn has several reasons.

1. **Speaker variability.** Individual people differ in terms of voice quality due to their differing physiology, differing states of health, and differing physical and emotional states. On a wider scale, they also differ in terms of their accent and language.

2. **Style variability.** The pronunciation of a word can differ markedly as a result of differing styles of speech, such as read speech, planned speech, or conversational speech.

3. **Environmental variability.** The physical environment in which the speech is uttered has an important effect on the recorded waveforms. Differing rooms lead to differing reverberant effects on the signal, and interfering noise by other speakers or the environment can have an overwhelming negative impact on the success of subsequent speech recognition.

In a DTW-based speech recogniser, such variability would have to be accounted for by storing multiple templates for the same word or sound. This quickly becomes impractical and computationally inefficient. A HMM can absorb this variability because it embodies a statistical description of the observation sequences. It is the current state-of-the-art practice to include as much training data as possible when determining the parameters of the HMMs, in the hope that the variability of the training material will be a good approximation of the variability the speech recogniser will face in practice. However, since a single model must account for an increasingly wide variety of signals, it becomes progressively less discriminative. While the performance of the system improves on average for a variety of speakers, speaking styles and environments, it will no longer do as well for any one particular case. This trade-off between *speaker-dependent* and *speaker-independent* speech recognition is well known. It is also the reason why systems which can reasonably expect to be used by a single or at least a small number of users offer *enrolment*: the process of adapting an initial speaker-independent set of acoustic models to the voice and style of a particular speaker.

Current state-of-the-art speech recognition systems are developed using several thousand hours of annotated speech, and accompanying hand-crafted pronunciation dictionaries. Such enormous resources represent a huge investment in terms of specialist linguistic expertise and time, and have as a consequence only been developed for a small proportion of the world's accents and languages. English systems, for example, may be available for North American and UK accents. South African languages and even South African English, on the other hand, have not yet been significantly accounted for. The scale of the task of compiling comparable resources for new language varieties is one of the chief obstacles to the development of speech technology.

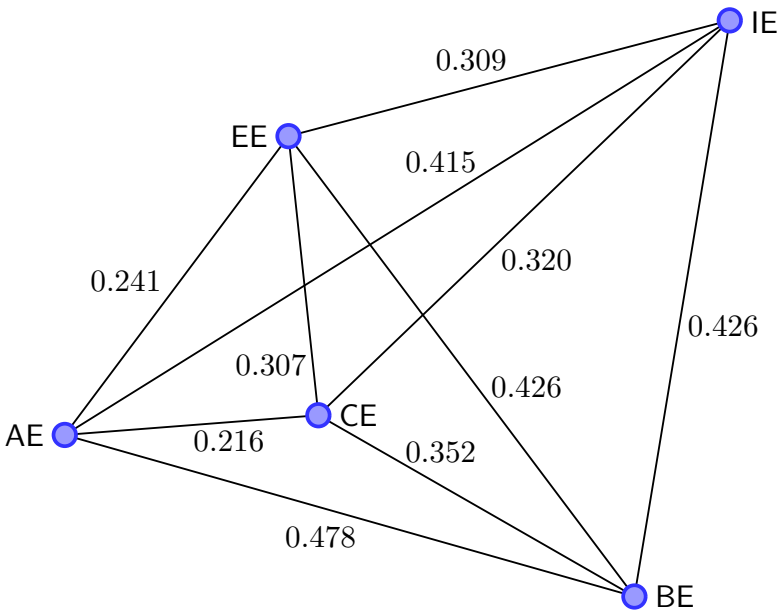

Figure 7: Statistical similarity between the five major accents of South African English: Afrikaans English (AE), Black South African English (BE), Cape Flats English (CE), White South African English (EE), and Indian South African English (IE)

Moreover, closer scrutiny reveals that labels such as 'US', 'UK' and 'SA' accents are rather imprecise, because in each case there is considerable regional variability. Particularly in South Africa, due to the large number of languages spoken within the same geographic region, this variability is considerable. Figure 7 illustrates the similarities between the five major accents of South African English recognised in the literature (Schneider et al., 2004). These distances were calculated by comparing, in a pairwise fashion, HMMs of corresponding phonemes for every accent pair (Kamper & Niesler, 2013). When the phonemes are, on average, more similar, the accents are located closer together in the figure. It is clear that the varieties of English spoken by Black and Indian South Africans differ markedly from the others. These differences lead to deteriorated speech recognition performance under mismatched conditions, for example when a Black speaker uses a system trained predominantly on speech by English mother-tongue speakers. Our own experiments have shown that for large vocabulary unconstrained English speech recognition, word error rates can more than double, rising from 28% to almost 60%.

Multi-accent speech recognition and the related problem of multilingual speech recognition are current subjects of research that are especially relevant in a multilingual society such as South Africa. Recent work has sought to find ways in which commonalities between languages and accents can be capitalised on, while distinctive aspects are modelled separately (Niesler, 2007; Kamper, Mukanya & Niesler, 2012). The two traditional approaches to the problem are either to model each language or accent separately (leading to language- or accent-dependent systems), or to pool all data and create a single set of HMMs that will model all languages or accents together (leading to language- or accent-independent systems). A more refined approach is to share data where this is beneficial, and to keep it separate where it is not. Our own work has shown that this does indeed lead to improved performance. However, the same advance had led to the renewed insight that speech characteristics, such as accent, resist the hard classifications that make dealing with them tractable from an engineering perspective. For example, it was discovered that the strength of an accent can vary for the same speaker, depending on whether the discourse is monolingual or not.

When more than one language occurs within the same dialogue, the speakers are said to be engaging in code-switching. This occurs particularly frequently among Black South Africans, who habitually alternate between their mother tongue and English. Our research has shown that the accent of the English used by these speakers varies depending on whether it is part of a mixed code or not. Figure 8 shows that, when isiXhosa and isiZulu speakers switch to English from their respective mother tongues, their English accent can be more accurately determined than when they are engaged in a monolingual English conversation (Niesler & De Wet, 2009). When code switching does not occur, the accent can be determined with an accuracy that is close to chance, indicating that the speech is not clearly accented. This is true when the judge is an automatic accent identifier, and also when mother-tongue human judges are questioned. Such continuous accent shifts present challenges for ASR that have not yet been addressed, and that are not encountered in predominantly monolingual societies.

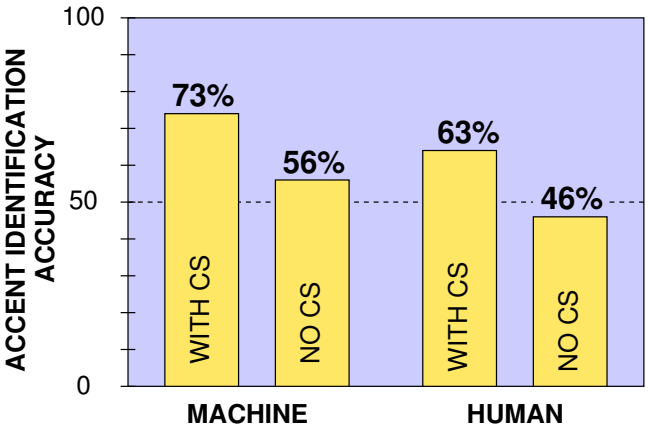

Figure 8: Accent identification accuracy for English spoken by isiXhosa and isiZulu speakers when the English is drawn from an utterance with and without code switching (CS). Identification accuracies are shown both for an automatic system (left) and for a perceptual experiment using human mother-tongue isiXhosa and isiZulu speakers (right).

6 CONCLUSION

The field of automatic speech recognition has developed steadily since its inception in the early 1960s. Speech recognition is now accessible to anyone with a smartphone and an internet connection, and so, in a sense, it has reached the mass market. However, this technology, which promises to liberate populations by enabling access to information without prerequisite computer literacy,

continues to fall short of its ideal. Deviations from the intended language, accent and means of discourse regularly lead current systems to fail. In part this may be ascribed to the predominance of a few language varieties with wealthy populations of speakers in the development of speech technology. But it is also due to the complexity of human verbal communication, which is still imperfectly understood. While advances in the neurological and psychological sciences will improve our understanding of the mechanisms behind such communication, the engineering sciences can address the limited reach of the systems by making them practical for a wider group of users.

My own research has become focused on methods that would advance speech technology in the South African context. This includes the development of multilingual and multi-accent speech recognition systems. Such systems are able to automatically process speech in a variety of input languages and accents, as is necessary for a highly multilingual society. A particular current challenge is the recognition of speech that includes code switching, since this necessitates the speech recogniser to alternate between languages within the same utterance. I am also engaged in the development of techniques that aim to reduce the investment that is currently needed for the development of speech corpora and pronunciation dictionaries to achieve competitive ASR performance. Because the development of these resources requires specialised skills and is exceptionally time consuming, it is very costly and a major obstacle to the development of speech technology for a new language or accent. Data-driven and self-organising approaches are being developed to reduce or even eliminate the human effort currently required. While such methods are extremely computationally intensive, they offer a realistic alternative to the development of speech recognition systems for the under-developed language varieties of Southern Africa.

REFERENCES

Baum LE, Petrie T, Soules G & Weiss N. A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *The Annals of Mathematical Statistics*, 41(1):164–171, 1970.

Bellman RE & Dreyfus SE. *Applied Dynamic Programming*, Volume 7962. Princeton University Press, Princeton, N.J., 1966.

Davis KH, Biddulph R & Balashek S. Automatic recognition of spoken digits. *The Journal of the Acoustical Society of America*, 24:637, 1952.

Kamper H & Niesler T. The impact of accent identification errors on speech recognition of South African English. *South African Journal of Science*, in press, 2013.

Kamper H, Mukanya FJM & Niesler T. Multi-accent acoustic modelling of South African English. *Speech Communication*, 54(6):801–813, 2012.

Markov A. Extension of the limit theorems of probability theory to a sum of variables connected in a chain. In R. Howard (ed.), *Dynamic probabilistic systems (Volume I: Markov models)*, Appendix B, pp. 552–577. John Wiley, New York, 1971.

Niesler T. Language-dependent state clustering for multilingual acoustic modelling. *Speech Communication*, 49(6):453–463, 2007.

Niesler T & De Wet F. The effect of code-mixing on accent identification accuracy. *Computer Speech and Language*, 23(4):435–443, 2009.

Rabiner LR. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.

Sakoe H & Chiba S. A dynamic programming approach to continuous speech recognition. *Proceedings of the Seventh International Congress on Acoustics*, 3: 65–69, 1971.

Sakoe H & Chiba S. Dynamic programming algorithm optimization for spoken word recognition. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 26(1):43–49, 1978.

Schneider EW, Burridge K, Kortmann B, Mesthrie R & Upton C (eds). *A handbook of varieties of English.* Mouton de Gruyter, Berlin, 2004.

Stratonovich BL. Conditional Markov processes. *Theory of Probability & Its Applications*, 5(2):156–178, 1960.

Vintsyuk TK. Speech discrimination by dynamic programming. *Cybernetics and Systems Analysis*, 4(1):52–57, 1968.

Viterbi A. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Transactions on Information Theory*, 13(2):260–269, 1967.