



HTK Limited
Hubbard House
6 Civic Drive
Ipswich IP1 2QA UK

Tel: +44 (0) 870 600 2311

Fax: +44 (0) 870 600 2312

www.htk.co.uk

HTK White Paper:

Language Identification Technology and Application

Issue 1.0

17th July 2007

HTK Company Confidential

© HTK Limited

Author: Denis Johnston

Contact: Marlon Bowser
0870 600 2311

Ref: PPD/P010/LIDTechnologyWhitePaper

Confidentiality Statement

Disclosure of this document or any part thereof without permission from HTK Limited is prohibited, and the reader is cautioned that such disclosure may cause harm to HTK Limited's business activities.

NOT PROTECTIVELY MARKED

Contents

Executive Summary.....	3
1. Introduction.....	4
2. The Nature of Language.....	4
3. Languages within the UK.....	5
4. Language Opportunities within the UK.....	7
5. Opportunities for New Enterprises.....	9
6. The Virtual Public Office.....	9
7. Language Identification Technology and Performance.....	10
8. The Underlying Technology.....	11
9. Alternatives and Complementary Approaches.....	12
9.1. Co-Operative Approaches.....	12
9.1.1. EARCON.....	13
9.2. Totally Transparent Options.....	13
10. Strategic Thoughts for Technology Suppliers.....	14
11. Conclusions and Recommendations.....	15
Technical Appendix: Identifying Language from the Speech signal.....	16
1. Introduction.....	16
2. Early Attempts at Automatic Language Identification.....	16
3. Modern Systems Based on Phonotactics.....	17
4. Full Recognition Systems.....	18
5. Recent Trends and Future Prospects.....	19
6. References.....	20
Revision History.....	20

NOT PROTECTIVELY MARKED

Executive Summary

Automatic Language Identification (LID) is a relatively unknown and untapped technology that has the potential to dramatically improve social inclusion in the public sector and commercial success in the private sector.

According to a DfES (then DFEE) report in 2001, there are an estimated 1 to 1.5 million people living in the UK that lack the English language skills required to function in society and employment. Just over 30% of the population of London was born abroad, and over 300 different languages are spoken in the Capital. The Nuffield Languages Inquiry (1998-2000) recommended that steps should be taken to ensure “no-one is denied access to essential public services by barriers of language”. Whilst many departments and local authorities have introduced multi-lingual websites, 85% of users access these services by telephone.

The richness of languages within the UK can however be viewed as a great asset. In a number of industries, especially those based on ‘trust’ such as broadcasting, insurance and banking, the UK has become a ‘language hub’. Communicating in the customer’s language can pay significant dividends:

“Limited English speaking consumers are four times as likely to buy a product or service in their preferred language. More and more companies are realizing that providing multi-language channels is not only a convenience for customers but a necessity for growing business in an increasingly diverse marketplace.”

(Jody Garcia, vice president of Specialty Customer Care for AT&T West.)

Reinforcing this view is the London Mayor’s Office ‘Diversity works for London’ initiative and ‘Diversity Dividend’:

“The range of languages spoken in the capital makes it an attractive location for global business”.

One route to realizing this growth may be through the “Virtual Public Office” (VPO) and a “One Contact – One London” service. If an effective form of call-steering based on language can be developed, radically different, even disruptive, ways of working can be envisaged.

Analysis of the required technologies suggests that such a service is technically feasible and would be suited to a centrally managed shared service architecture. It would require a combination of cutting-edge speech recognition along with more transparent methods such as CLI and user-preference based systems. Looking forwards, the industry standardisation onto IP-based networks such as BT 21CN will simplify integration with other services, and enable language identification to be deployed as part of a joined-up and ‘customer-centric’ strategy.

The challenge lies in applying these technologies and the associated human-factors psychology effectively, and the opportunity for ‘first-mover’ advantage is immense. Not only will first-movers gain intellectual property and service differentiation benefits, but by incorporating Language Identification Technologies they will facilitate far greater social inclusion, extend their customer base and achieve international recognition as a centre of excellence.

HTK is a market leader in the application of LID and can provide the necessary expertise to support organisations looking to invest in LID technology.

NOT PROTECTIVELY MARKED

1. Introduction

Automatic Language Identification (LID) has the potential to extend the customer base of existing Interactive Services and provide 'language call-steering' for a number of socially important and commercially novel applications.

This white paper highlights existing problems that could be directly solved using managed services incorporating LID technologies, and looks to areas where the same approach could offer major growth opportunities for a number of UK industries.

It provides a summary of this relatively unknown and untapped technology and shows that LID is not only ideally poised to solve a number of 'unmet needs' in telephony applications, but has major commercial potential.

It begins with some remarks concerning the nature of language. It then describes how the linguistic diversity within the UK – and especially within London – not only provides a number of social challenges which LID may help to solve, but shows that this diversity can be harnessed in emerging markets.

This is followed by a section in which the LID technologies, their strengths, limitations and performance, are addressed.

Finally, it demonstrates how a number of current problems can be solved and future opportunities realized by the imaginative application of a range of call-steering techniques.

2. The Nature of Language

Estimates of how and when language originated has divided scholars for centuries. Until recently it was thought that language originated only 40,000 years ago, but the latest paleoanthropological evidence has revised this to over 100,000 years.

The latest edition of Grimes' "Ethnologue" catalogues 6,912 separate languages. Population figures are available for just over 6,000 of them.

Of these 6,000;

52% are spoken by fewer than 10,000 people;

28% by fewer than 1,000;

83% are restricted to single countries;

Just over 2,000 have a written form.

In many instances, deciding whether languages are 'different' is far from straightforward, as people who speak one (e.g. Danish) can often understand others (e.g. Swedish and Norwegian).

Variations *within* a language can sometimes be greater than those *between* languages – especially where there may be different 'high' and 'low' versions or even (as in Japanese) where different politeness levels co-exist. Some may be similar in the spoken form but use different written scripts. For example, with a little practice, Polish and Ukrainian speakers can "tune-in" and understand each other, but they systematically use different word endings and the languages have entirely different scripts.

NOT PROTECTIVELY MARKED

Also, languages continually evolve; absorbing and losing words and phrases.

Entirely new “constructed languages” may also be invented. Klingon, for example, is a completely contrived language for the television series StarTrek. Tolkien, author of ‘Lord of the Rings’, created his own series of languages at least two of which, Quenya and Sindarin, had extensive vocabularies and grammars.

There is evidence that new languages emerge and evolve extremely rapidly, and this happens especially in isolated groups. Even religious or otherwise tightly bonded groups will often generate their own languages, or radically modify their first language so that it distinguishes that group to be ‘different’.

Deaf children in particular, left to their own devices, will spontaneously create their own sign languages. In a part of Ghana where, due to hereditary factors, there is a higher than usual level of deafness, the Adamorobe language (which is primarily a sign language) has evolved. Although it is estimated that only between 2% and 15% of the population are actually deaf, this sign language is the primary means of ‘spoken’ communication in the whole community.

3. Languages within the UK

Half a century ago, the only non-English language spoken by any significant numbers of the UK population was Welsh (with over 500,000 speakers in 1960). Very few were monoglot Welsh speakers, and now there are probably none. For most people in the UK, the closest they got to practical experience of another language was reading French on the back of the label on a bottle of HP sauce!

According to the most recent National census data (for 2001), 8.3% of the population was foreign-born and according to the same census, the largest ethnic groups for people born overseas were Indian (569,800) and Pakistani (336,400).

The National Centre for Languages currently estimates that over 700,000 UK children are bilingual, with over 300 languages covered.

However the figures for languages spoken are much more difficult to obtain and open to interpretation.

The list below provides some estimates of languages and the numbers of speakers.

Assyrian (5,000), Bengali / Sylheti (400,000), Cornish (100), Eastern Panjabi (471,000), Estonian (14,000), Greek (200,000), Gujarati (140,000), Hakka Chinese (10,000), Gaelic / Irish (74,000), Gaelic / Scottish (58,000), Hebrew (8,000), Hindi (243), Italian (200,000), Japanese (12,000), Kashmiri (115,000), Latvian (12,000), Leeward Caribbean Creole English, Malayalam (21,000), Maltese (40,900), Mandarin Chinese (12,000), Mirpur Panjabi (20,000), Moroccan Spoken Arabic (5,800), Northern Kurdish (23,766), Parsi (75,000), Polish (700,000), Portuguese (17,000), Shelta (30,000), Sindhi (25,000), Somali (1,600), Southern Pashto (87,000), Southwestern Caribbean Creole English (170,000), Tagalog (74,000), Ta'izzi-Adeni Spoken Arabic (29,000), Tamil, Turkish (60,000), Urdu (400,000), Vietnamese (22,000), Welsh (650,000), Western Farsi (12,000), Western Panjabi (102,500), Yoruba (12,000), Yue / Cantonese Chinese (300,000)

(Sources include Ethnologue 2005, Government statistics, National Centre for Languages CILT, BBC Immigration maps and the Guardian Ethnicity surveys.)

NOT PROTECTIVELY MARKED

Unfortunately, knowledge of the country of birth or the fact that people can speak a language other than English, does not really help in providing a useful estimate of English language competence. More importantly, it does not indicate the extent of difficulties that may arise for non-English speakers either in work or when dealing with public services.

In 2001, the Department for Education and Skills (then DfEE) commissioned a report to examine the needs of people whose first language is not English. It focused on barriers to employment, education and training.

However the first of the key findings from “English language as a barrier to employment education and training” states that;

“There is no reliable data on the number of people living in Great Britain whose first language is not English. This causes serious problems with the planning and delivery of education and training provision.”

However it goes on to say;

“At least three million people living in the United Kingdom were born in countries where English is not the national language.”

And concludes that;

“An estimated 1 to 1.5 million people living in the UK lack the English language skills required to function in society and employment.”

Many of these people will of course be looked after by relatives, friends and communities who can provide the necessary interfaces, however these numbers suggest that there is a significant wasted potential.

The Nuffield Languages Inquiry (1998-2000) was set up primarily to address the question “Where are we going with Languages?” Although, in the main it was concerned with language teaching and the needs of business, it also touched on the social dimensions.

Specifically, it said that;

“Language is fundamental to equality of opportunity and there is much to be done if people whose first language is not English are to exercise their right in the UK of informed access to services such as healthcare, justice, employment advice, housing and social services. Investment in the domain of language technologies will assist in providing solutions to some of the issues.”

Adding;

“Public services: The role of languages. Fluency in English can be deceptive. The UK has always been a multilingual society, and many people whose first language is Urdu, Hindi, Chinese or Turkish may also be fluent speakers of English. Even so, public services can still present obstacles to people from other language backgrounds because, despite being fluent in English, they may feel more in control when interacting in their first language. For non-speakers of English trying to access and negotiate public services, the obstacles can be insurmountable without outside help.”

NOT PROTECTIVELY MARKED

As a result, one of the recommendations was;

”Recommendation 1.9: Ensure that no-one is denied access to essential public services by barriers of language”.

Although not a statutory requirement, most local authorities and government agencies are now aware of this need. Many now produce multilingual documents and have multilingual websites. However, 85% of service users access services via the telephone and it is here where the major weaknesses are believed to lie.

But it is the emergency services ‘on target’ statistics that indirectly provides evidence that a deeper problem exists. One of the established causes of ‘missed targets’ in the ambulance service is the delays which arise when a non-English speaker calls.

Although there are various strategies for dealing with this, they are all extremely slow – basically passing the call on to someone who might be able to understand. This ‘call-steering’ is an area where there is a very clear ‘unmet need’ and a clear political will for improvement, and one where any improvement would have a significant quantifiable benefit.

With the advent of the Single Non-Emergency number (SNEN) the need for call-steering can only increase, and the pressure to reduce costs will be correspondingly greater.

Although not so critical, but on a much larger scale, Local Authorities and Government Departments experience similar difficulties. The problem here is not so much a lack of interpreters, or other officers who understand the language. Rather, the problem is that of initial language identification and the logistics of rapidly finding and connecting the call to those with the essential language skills.

4. Language Opportunities within the UK

While all of this might seem to suggest that UK language diversity is nothing more than a source of problems, recent thinking suggests that this is very far from the truth.

The richness of languages within the UK means that in a number of industries the UK has become a ‘language hub’. Especially in industries based on ‘trust’, such as broadcasting, insurance and banking, communicating in the customer’s language can pay significant dividends.

More than ever, language now matters.

That is why, far from considering UK language diversity as a problem, it is starting to be viewed as a major national asset.

This is especially the case in London where diversity is greatest. The London Mayor’s Office has several initiatives including “Diversity works for London” along with the idea of the “Diversity Dividend”;

THE DIVERSITY DIVIDEND

“The range of languages spoken in the capital makes it an attractive location for global business”.

NOT PROTECTIVELY MARKED

“Companies can locate in London in the knowledge that they will be able to access people with hundreds of languages, communicating with the rest of the world from one city.”

“London's diversity is its greatest strength and was crucial in winning the race to host the Olympic and Paralympics Games in 2012. I want to ensure that London businesses really harness and make the most of the opportunities that diversity in the workplace can deliver. “

(Mayor of London Ken Livingstone)

And, along with business is tourism – worth £14.2 billion from visitors in 2006. However, the citizens of the UK spent £30 billion abroad as tourists – more than twice as much.

This seems to be an area where the UK could do better, and it was an area singled out for comment in the Nuffield Languages Report which put it bluntly;

WELCOMING VISITORS AS LONG AS THEY SPEAK OUR LANGUAGE

“Two thirds of the 25 million tourists who visit the UK each year come from countries whose first language is not English, yet our travel and tourism industry still depends largely on its customers taking the trouble to learn our language. This is a scandalous picture....”

(Nuffield Languages Report)

Of course many of these visitors speak English as a second language, and indeed many will come simply in order to be ‘immersed’ in English.

But that simply begs the question of how many *more* – and from the fastest growing regions such as China – would come if they could get along better in their own language. In a global marketplace, anything that makes things just a little bit easier for the customer will always pay dividends.

“Today as never before we are having to search out new markets, often where there is little tradition of speaking English, and where we hope to increase our trade substantially. We are simply not going to prosper without mastering International Communication Strategies.”

(Brian Wilson MP - Prime Minister's Special Representative on Overseas Trade)

And;

“Other Nationals do not learn English for our benefit!”

(Nuffield Report)

Important as they are, it looks as if the markets serving social needs and the tourist industry could be dwarfed by other commercial applications of language technology.

There are now signs that the opportunities for LID go far beyond these ‘obvious’ industries. Perhaps unsurprisingly, the agent of change is “globalization”.

NOT PROTECTIVELY MARKED

5. Opportunities for New Enterprises

A tantalizing glimpse of this potential comes from AT&T. AT&T has traditionally used systems such as Language Line for basic emergency and help services. However, it has recently discovered that by offering to respond to a wide range of inquiries in the customer's own language it gives them a major sales advantage. Customers are up to *four times* as likely to buy a product if contacted in their preferred language;

"Limited English speaking consumers are four times as likely to buy a product or service in their preferred language. More and more companies are realizing that providing multi-language channels is not only a convenience for customers but a necessity for growing business in an increasingly diverse marketplace."

(Jody Garcia, vice president of Specialty Customer Care for AT&T West.)

The reasons are self evident. They are drawing-in a large number of customers who – although they could speak English – were not sufficiently confident to undertake a telephone transaction.

That lesson should not be lost upon UK retailers.

With increasing globalisation, "language friendly" services will become the norm, and Language Identification Technology will become a necessary first step for anyone who hopes to grow their services within, and beyond, the UK.

6. The Virtual Public Office

One route to realizing that growth may be through the "Virtual Public Office" (VPO). If an effective form of call-steering based on language can be developed, radically different, even disruptive, ways of working can be envisaged. If the language capability and preferences of the caller can be identified, and if calls can be seamlessly routed to agents available to speak fluently in the caller's preferred language, the potential customer base (and customer satisfaction) of any enterprise is immediately increased.

In the shorter term there is an even more urgent need if the "One Contact – One London" theme is to be realized. Here, a single number will provide the access point for all of the major information services in London, and if it is to be truly 'universal' it will have to accommodate the needs of many non-English speakers.

And then there is the Olympics. More than anything, it could be said that it was the 'diversity dividend' that brought the Olympics to London.

With those games now less than 5 years away, the need to become *the* 'language friendly' nation of the world is not just an option – it is a major imperative – and one which provides a unique opportunity to showcase Innovative UK technology.

But what exactly are the technology options? How well does the technology work? Is it cost effective and what needs to be done?

The next section addresses these questions.

NOT PROTECTIVELY MARKED

7. Language Identification Technology and Performance

Almost all advanced research and development in Language Identification concentrates on tackling 'freely spoken speech'. This is not just because it is a particularly interesting academic challenge but also, no doubt, because the only perceived useful application for LID was the clandestine identification of signals by the National Security Services.

In recent years, much of this research has been steered by NIST (the National Institute for Standards in Technology). Based in the USA, but open to worldwide participation, NIST does not carry out the research itself but provides a 'competitive framework' which encourages disparate groups to work, so that their results can be meaningfully compared.

NIST normally brings the various participants together every-other year at a workshop where the various techniques are compared and contrasted. The last workshop was in 2005 and another is due in August 2007.

In the case of Language Identification, the main technical role of role of NIST is to specify test material in the form of an 'unseen' database of audio recordings and to define the test procedures.

In a typical test, a system designed to identify a single 'target' language is subjected to several thousands of test utterances. Some of these utterances are in the target language but most are fragments of speech from other languages. A perfect system, making no errors, would accept as valid all speech in the target language and reject all utterances spoken in any other language.

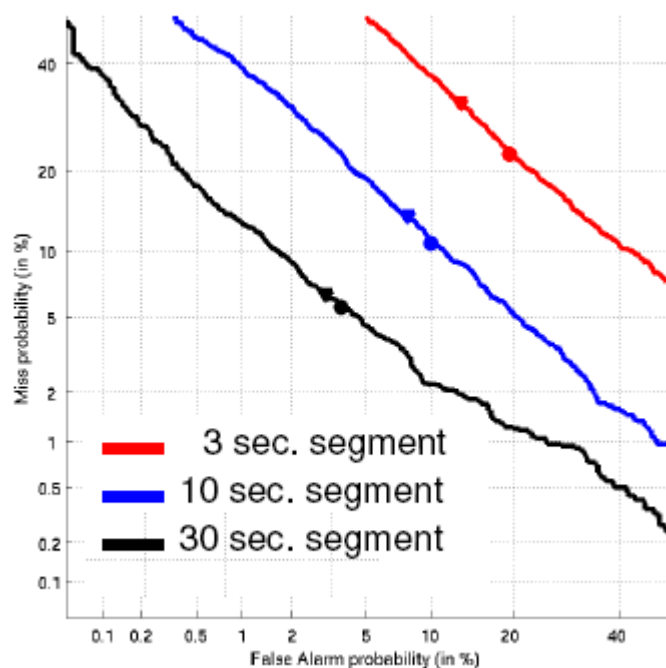
In practice no system is perfect, so individual systems are rated according to the percentage of 'target language' fragments that they fail to detect (misses) and the percentage of 'other language' speech fragments they incorrectly identify as the target language (false alarms).

A complication arises here because not only are there two types of errors which are important - misses and false alarms - but it is always possible to adjust the system to trade these factors off against each other. To deal with this unavoidable problem, tests are made using several different settings of decision thresholds and the results are plotted as graphs showing the error rates observed over the range.

Such curves are known as error rate detection (DET) curves.

In the 2005 evaluation, 12 laboratories took part and 7 languages were involved (Mandarin, English, Japanese, Korean, Spanish, Tamil and Hindi). A summary of the results is available in a slide set (A.F. Martin, A.N. Le: "The current state of Language Recognition - NIST 2005 evaluation Results") which may be downloaded from <http://www.speakerodyssey.com/templates/56.pdf>

The DET curves shown there, one of which is reproduced below, provides an indication of the performance which can be expected from today's leading systems. The vertical axis represents 'true target misses', and the horizontal axis represents the 'false alarms'.

NOT PROTECTIVELY MARKED

A useful single figure indicator of quality is given by the equal error rate (EER). This is the error rate where the proportion of misses equals the proportion of false alarms. The data shown here suggest Equal Error Rates (EER) of 5%, 10% and 20% for speech segments of 30, 10 and 3 seconds respectively.

The latest NIST summary shows that there is a wide difference between languages, with some being identified with a probability of error of only 2% EER, while others have a probability of error as high as 20% for the 30 seconds segments.

8. The Underlying Technology

Many of the underlying techniques of LID are similar to those used in Automatic Speech Recognition (ASR). A single ASR system working in one language tends to be computationally expensive, and as LID systems will have to simultaneously process every speech fragment through tens or even hundreds of parallel recognition systems, each tuned to a different language, their computational load is usually proportionately much greater.

To increase efficiency and reduce this computational load, most current LID systems first try to extract features at a common phonetic level and then examine the phonotactic patterns (the ways in which the phonetic features are combined) to see how well they match known distributions within the target languages. This can be undertaken with varying degrees of sophistication, from the simplest that simply count phonotactic coincidences, to those that try to reconcile the sounds with all that is known about all of the target languages.

For some time it has been generally accepted that the more 'knowledge' about a language that can be built-in to the recognition process, the better the performance. However the most recent empirical evidence, especially from the NIST 2003 and NIST 2005 results, suggests that despite massively increasing complexity, only marginal improvements in performance are now being made. It is unclear if

NOT PROTECTIVELY MARKED

this means that a natural limit of performance has been reached, but strongly suggests that short term evolutionary improvements are unlikely.

The computing power requirements, the need for regular updating of the language models, and the fact that the demand for such a service will tend to come from a widely dispersed user base, means that a centrally managed voice service is likely to be the preferred implementation.

Such a service is technically feasible and would provide a flexible, generic solution.

However, not all applications may require such sophistication. In some circumstances much simpler solutions may be sufficient, and these are explored next. Combining the different approaches may provide the optimal solution from aspects of cost, usability, accuracy and other perceived benefits.

9. Alternatives and Complementary Approaches

The techniques just described tackle the problem head-on and attempt to identify languages in all their richness and variety.

But in real situations, such as emergency calls, what people say will usually be quite constrained. Their needs will be focused and dominated by words and phrases relevant to the problem. Words such as "Help", "Quickly", "Fire", "Police", "Ambulance" and phrases such as "I don't speak English" (although spoken in the caller's language) are likely to occur very frequently and will often be repeated within the call. Even if the vocabularies used extend to hundreds or even thousands of words and phrases, this is a tiny fraction of the total number of words in a language and thus offers significant scope for simplification and improvements in overall system accuracy.

Unfortunately, there are no reports of such techniques ever having been attempted, let alone developed, but perhaps this may be taken as indication of the potential for application-focused innovation in this area.

Other alternative strategies tend to fall into one two of two categories;

- 1) Those that require some user co-operation (or lack of co-operation!) at the time of use.
- 2) Those that are, or become, completely automatic, i.e. "invisible" to the user but which may involve some enrolment or registration.

9.1. Co-Operative Approaches

One very simple 'co-operative' approach extends a technique already used in many local government offices. Telephone agents and operators are told to simply say "English?" when they hear non-English speech, and clients are expected to respond with the name of their language.

In many communities, especially where there is thought to be a potential problem, this protocol is widely advertised by means of mail-drops and flyers. Provided that the client knows what to do and the operator is able to recognise the name of the language, this provides a simple, friendly and effective means of steering the call to an appropriate interpreter.

NOT PROTECTIVELY MARKED

Clearly, there is a role here for an automated managed voice service using conventional ASR, as it may simply be necessary to recognize a single word.

In a rather bizarre twist, in 1996 the Japanese company KDD used a language identification system which relied upon a *lack* of a valid response. KDD had begun to experience an epidemic of ‘prank calls’ to users of its ‘Home Direct’ service. These were International calls originating from a South American country where school children had discovered and distributed the Japanese Home Direct free-phone number. The children were dialling this in their thousands, just to listen to the Japanese operators talking to them.

The KDD solution was to intercept all calls to the free-phone number and play a message in Japanese; “Today’s password is xxxx. Please say the password”. Only bona-fide Japanese speakers understood what had to be done, and a speech recogniser confirmed that they were responding appropriately. A very simple, but highly effective, solution.

Some estimates put the KDD savings in to the millions of pounds.

Others solutions are exemplified by TouchTone[®] based services. If a caller can simply type in a number or use the keypad to text-type the language, then this provides the user with a familiar and exceedingly easy way to announce that a service is required in a particular language.

The problem is that it is necessary to tell the user that this option is available, and this needs to be done in his or her language... A “catch 22” situation.

For those applications where the number of languages is small, then a simple dialogue-based solution is likely to be adequate; e.g. “Press 1 for Welsh and 2 for English” (repeated in Welsh). This may even be appropriate for four or even five languages, but above that number it becomes exceedingly tedious and is often considered an irritant by the vast majority of users who speak the dominant language.

9.1.1. EARCON

One intriguing solution, but one which has never progressed beyond the concept stage, is to establish a universal EARCON – a musical sequence which all telephony system users would recognise as an invitation to indicate their preferred language. On hearing this universally accepted tune, users would simply key-in their language code or speak the name of their preferred language.

However, any solution that requires the caller to understand and interact with the system will not only require some form of dialogue, but will inevitably introduce transaction delay.

Which begs the question: Are there solutions (or partial solutions) that are totally invisible to users?

9.2. Totally Transparent Options

One of the simplest options (albeit fraught with issues concerning privacy) is to use calling Line Identification (CLI) coupled to an exceptions database which holds a list of preferred languages for registered telephone lines.

A variant of such a solution is actually in widespread use already – but more by accident than design. The author recently had a train journey where a woman sitting across the aisle received a constant stream of cell-phone calls which she immediately answered in English, French and at least one other

NOT PROTECTIVELY MARKED

language that I couldn't identify. She was able to do this simply because the caller's name (or possibly International phone number) appeared on the cell-phone display, and of course she immediately knew the appropriate language to use in answering the call.

Slightly more complicated and also fraught with privacy issues, but easily within the compass of today's technology, is to use CLI coupled with "Reverse Directory Enquiries". As a person's name can often indicate their country of origin and hence language set, this may be used to 'short list' some candidate languages.

Anyone who has ever used a bank card to access cash from a "hole in the wall" abroad will be familiar with the fact that the machine will often present language choices. This is an example of "*language recognition by token*". In this case the bank card – the *token* – holds a code that indicates the language of the holder.

Such tokens are already widespread. SIM cards used in mobile phones hold details of the language to be displayed on the screen; indeed this is also often user selectable. Although this information cannot (currently) be obtained easily for application use within a voice call, there are no fundamental technical reasons why this kind of functionality could not be added. Indeed, within next-generation IP-based networks (e.g. BT 21CN and IMS / SIP architecture) this capability will almost certainly become more "open" if not commonplace.

While such technologies may be used for telephony, the really big opportunities may lie where they are coupled with other applications.

Already there are pilot schemes where passports and ID-cards are Radio Frequency ID (RFID) enabled. RFID-based tokens can also be hidden in jewelry, worn as badges or dog-tags, or even embedded under the skin. There are even suggestions that a new generation of semi-conducting inks would allow RFID circuitry to be tattooed onto the skin (it can already be ink-jet printed onto paper and other materials).

Once a person becomes 'RFID enabled' the opportunities become almost limitless. 'Personalised public transport systems' that detect a person's language via an RFID tag in the ticket; shops with electronic displays that change according to the language of the viewer; advertisements on escalators that adapt according to the language of the person approaching; adaptive Billboards; hotels and restaurants that provide details in the language of the visitor; travel displays on buses and trains to provide information in all the languages of people in a carriage at any particular time.

Viewed in that way, LID becomes an essential component of the premium personalised product and services market.

10. Strategic Thoughts for Technology Suppliers

Numerous government reports in the UK routinely recommend that councils, local services and business include 'language issues' as part of their strategic thinking.

But few do – in the main because they do not know what can be achieved.

There is therefore, on the face of it, a ready market for those who understand language and who can offer packaged support in language technologies.

NOT PROTECTIVELY MARKED

There is also a major 'first-mover' opportunity. Language technology may have the potential to be *disruptive* technology by changing public expectations of what is possible. In particular LID may enable 'self-service' language applications to develop.

A managed service offering spoken-language identification is one possibility, and a company that could establish a language EARCON both as part of its brand and as a universal symbol would be in a strong marketing position. In the advertising and marketing sectors, 'the power of the jingle' (think about Intel) is well known – even when it has no direct relevance to the product.

The significance of being a 'first mover' is also important from a performance perspective. Speech and language technologies tend to 'bootstrap' and system performance rapidly improves as it is tuned to real data. Within a short timeframe, the performance of established systems is much better than those that have not had the benefit of this learning phase. This makes the task of catching-up more difficult for those who were not so quick off the mark.

Another way in which LID could act disruptively is by "globalizing" the Virtual Office. This goes beyond offering locally-based language interpretation as an add-on to existing Virtual Office solutions. Routing calls across the planet is now so easy that one can imagine the linguistic equivalent of a 'VISA' service emerging, where a French speaking 'Virtual Office' user would automatically be routed to French centre, or a Kannada speaker to a Bangalore based centre.

11. Conclusions and Recommendations

Language Identification is a major untapped technology. Opportunities lie not just in the area of social services provision, but with imaginative thinking can be applied to major private-sector commercial opportunities.

London in particular, is possibly the most linguistically and culturally diverse city in the world. It is also probably unique in that this is both widely celebrated and actively encouraged - especially through initiatives from the Lord Mayor's Office.

It is this diversity, and the support for it, that is widely credited as having been the 'clincher' for the 2012 Olympic Games bid.

That being so, and with run up to that Olympics having now started, the message seems clear; if ever the time was ripe for a major initiative to develop multilingual solutions in the UK, it is today.

It is also clear that innovative technology solutions abound; the challenge lies in applying them and establishing them effectively.

The opportunity for 'first-mover' advantage is immense. Not only will first-movers gain intellectual property and service differentiation benefits, but by incorporating Language Identification Technologies they will facilitate far greater social inclusion, extend their customer base and achieve international recognition as a centre of excellence.

Technical Appendix: Identifying Language from the Speech signal

1. Introduction

Most of us speak one language very well. Some of us speak a second or even a third with equal fluency and a select few may speak up to a dozen languages. The record is currently held by Ziad Fازه who in 2006 was confirmed to be fluent in 56 languages.

Once we understand a language we are also extremely good at detecting it, and can identify it having heard only a few words; often in less than a second.

Perhaps more remarkably, we can even tell a speaker's first language simply by listening to them speak in our language. We can often tell the birth language of a French or German person speaking English, and can even make a good guess as to the area where a language comes from even if we cannot be precise; "it sounds like Scandinavian", for example. Almost all of us have been embarrassed attempting to speak a foreign language abroad, receiving an immediate response in English – illustrating that our own native accent still shines through!

The fact that we can identify 'known' languages so incredibly quickly, accurately and effortlessly proves that sufficient information is available within the first few seconds of an utterance. So, surely automatic language identification (LID) should be extremely easy.

This paper discusses that challenge, and the achievements made to date.

2. Early Attempts at Automatic Language Identification

Compared to research into spoken word recognition, Language Identification was comparatively late on the scene. One of the first documented attempts was reported by Leonard and Doddington [1].

They describe experiments in which a number of passages of speech from five languages were transcribed using the phonetic symbols used by linguists to describe the individual sounds of speech.

The frequencies of these symbols and the probabilities of "symbol chains" in each language were then estimated. Tests were then carried out to see if the language could be identified using these elementary statistics. The results are reported as having been "...fairly encouraging on three of the five languages. However the other two languages were poorly recognized indicating a need for further work."

/ˈspɔʊkən/ /læŋgwɪdʒ/

The words 'spoken language' transcribed into the International Phonetic Alphabet (IPA)

In 1997, again with hand-labelled phonetic transcriptions, House and E. P. Neuberg [2] used Hidden Markov Models (HMM's) to capture the sequences of phonetic symbols. Using this very powerful mathematical technique, they claimed to show that 100% accuracy could be obtained on eight languages. Unfortunately, they used part of the training material as the test-set so these results tell little about how the system would have performed on 'real' data.

NOT PROTECTIVELY MARKED

In 1986, Foil [3] described a whole series of results using 10 seconds of speech, processed in a number of ways. He used 45 different types of vocal features (pitch, spectrally-based features, energy etc.) and all with different types of classifier. Training was generally conducted on several hours-worth of speech, with testing performed on a few seconds.

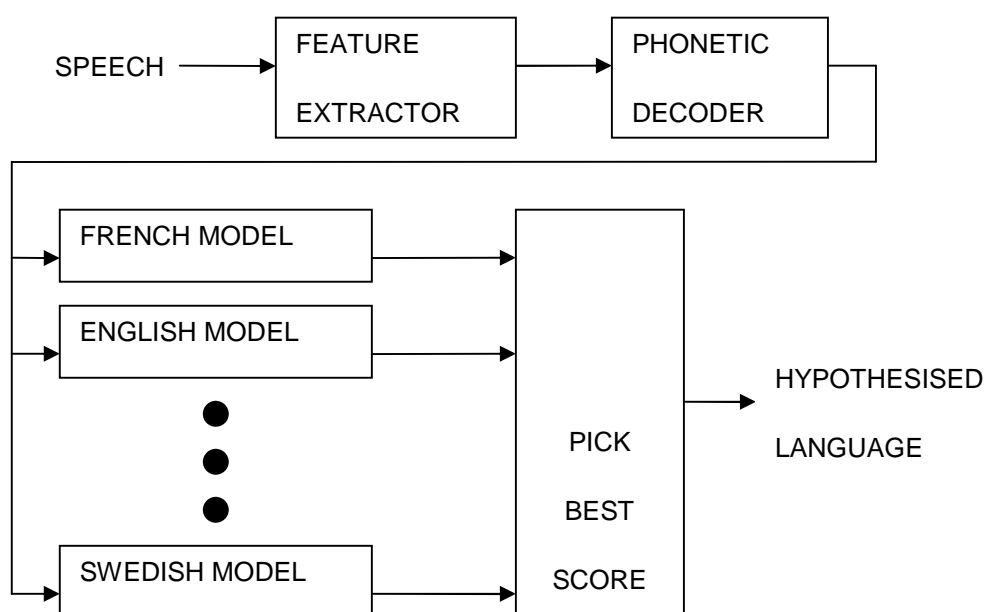
However, although the results were excellent where forms of human labelling were used, they were very disappointing for the fully automated systems. Here, the highest accuracy reported for three languages was 64% (corresponding to an error rate of 36%).

What all of these results seemed to confirm was that *if* the phonemes can be accurately determined, the phonotactics (the way in which the phonemes can be sequentially combined) is sufficient to distinguish languages. However, there was not yet an automatic way to detect the phonemes.

The great attraction of such an approach is its simplicity. If speech can simply be converted into strings of phonemes, or tokens that are in some ways similar to phonemes, then all that is required to build a language identification system is a database of audio samples of that language – or better still a database with phonetically labeled data. This avoids the need to have grammar models, dictionaries and all the complex paraphernalia of Automatic Speech Recognition (ASR) systems. It also means that a common ‘front end’ processor can be used for all languages; greatly reducing the computational load.

3. Modern Systems Based on Phonotactics

Modern Phonotactic-based systems typically use ‘front ends’ that are similar to those used in ASR systems. Here, the speech is typically chopped-up into overlapping frames of a few tens of milliseconds. Each frame is further processed to extract features – typically ‘industry standard’ Mel Frequency Cepstral Coefficients (MFCC’s).



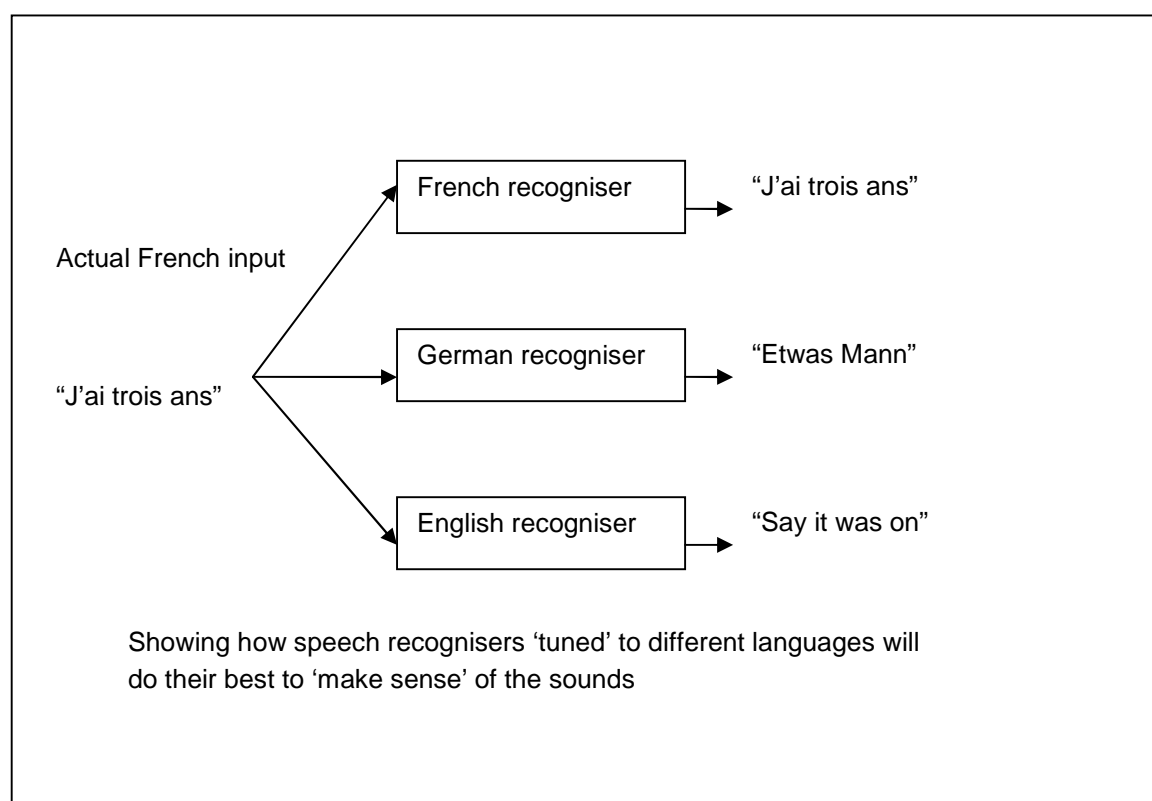
NOT PROTECTIVELY MARKED

The next stage is that of ‘tokenisation’. This is where LID diverges from the standard approaches of ASR. The tokenisation stage involves examining sequences of frames and clustering them in such a way that they can be labeled or indexed before passing them to a “pattern matcher”. The pattern matcher then determines the language by comparing the statistics of sequences to those of the known languages.

In practice, a separate tokenization process is used for each language – reflecting the belief that all languages tend to have different phoneme inventories – and this strategy does produce better results.

4. Full Recognition Systems

Another way of performing LID might simply be to use a number of normal ASR machines in parallel, with each ‘tuned’ to a target language. The machine that gave a sensible output (as opposed to gibberish) would be indicative of the language.



The direct application of this does not work well, simply because ASR systems have built-in language models that constrain them to output valid words in a ‘correct’ grammatical structure. However, a variant on the approach in which the ‘quality of fit’ is used can be very successful.

The first results obtained from these systems appeared in 1994. Papers by Tucker and Carey [4] reported 90% accuracy on a three-language test (English, Dutch and Norwegian).

This rose to 97% when just English and Dutch were compared. At the same time, Zissman [5] reported similar figures for a combination of English, Japanese and Spanish.

NOT PROTECTIVELY MARKED

While complex and computationally expensive, these techniques use the vast amounts of statistical information that model not just the sounds of language, but the ways in which they are combined into words and sentences.

However, they are currently limited in application because the necessary linguistic coverage is only available for a small number of languages.

5. Recent Trends and Future Prospects

Recently there has been a trend toward combining the output of both the simpler phonotactically-based systems and the fully blown ASR-based systems. This does appear to give improved results, which implies that the two systems are extracting independent information.

However, the fact that researchers are trying to squeeze the last drop of performance by marrying systems is symptomatic of the fact that, in recent years, the performance of each has levelled-out; remaining far below that of humans. This has been particularly noticeable in the NIST [6] tests of 2003 and 2005, where it appears that little basic improvement has emerged between the two dates, despite a massive increase in computational complexity and data refinement.

This suggests that any evolutionary improvements will be relatively slow and / or expensive, but it has also been the case that LID has been an area with very little active research in recent times. Virtually no papers were published in 2000 and 2001, but numbers are now increasing – much of it in the aftermath of 9/11.

Although the complex ASR-based methods have established themselves over the past decade, in the past couple of years the older phonotactic approach has caught up somewhat. One variant, known as the 'bag of sounds' by Lim and Li [7], and which incorporates a number of very novel approaches, reports extremely low error rates – less than 1% on a set of five Asian languages using samples of less than 15 seconds.

This has still to be ratified by a direct NIST-style 'head to head' comparison, but it does show that there are still some novel ideas to be explored, and progress to be made.

NOT PROTECTIVELY MARKED

6. References

- [1] R.G. Leonard and G.R.Doddington "Automatic language identification", Technical report RADC-TR-74-200, Air Force Rome Air Development Center, August 1974.
- [2] A. S. House and E. P. Neuberg in "Toward automatic identification of the language of an utterance. I. Preliminary methodological considerations", (Journal of the Acoustical Society of America, 62(3):708-713, 1977.)
- [3] Foil J "Language identification using noisy speech Proc ICASSSP 1986 pp861-865.
- [4] R. C .F. Tucker, M. J. Carey, E. S. Parris, "Automatic Language Identification Using Sub-Word Models", in Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing 94, Adelaide, Australia, April 1994.
- [5] M. A. Zissman, E. Singer, "Automatic Language Identification of Telephone Speech Messages Using Phoneme Recognition and N-Gram Modeling", in Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing 94, Vol. 1, pp.305-308, Adelaide, Australia, April 1994.
- [6] Martin A.F., Le A.N: "The Current state of Language Recognition NIST 2005 Evaluation" downloadable from <http://www.speakerodyssey.com/templates/56.pdf>
- [7] Boon Pang Lim; Haizhou Li; Bin Ma, "Using Local & Global Phonotactic Features in Chinese Dialect Identification" Acoustics, Speech, and Signal Processing, 2005. Proceedings. (ICASSP Volume 1, Issue, March 18-23, 2005 Page(s): 577 – 580.

Revision History

Issue	Date	Author	Scope of Changes
Draft 0.0.1	6 Jun 2007	D Johnston	First draft for internal review.
Draft 0.0.2	26 Jun 2007	D Johnston	Second draft for internal review.
Draft 0.0.3	6 Jul 2007	D Johnston	Includes technical appendix and exec summary.
Draft 0.0.4	13 Jul 2007	D Johnston	Correction to various typos and general tidy-up.
Draft 0.0.5	16 Jul 2007	D Johnston	Final draft for review and sign-off.
Issue 1.0	17 Jul 2007	D Johnston	Issued.