AN EXAMINATION OF THREE CLASSES OF ASR DIALOGUE SYSTEMS: PC-BASED DICTATION, IN-CAR SYSTEMS AND AUTOMATED DIRECTORY ASSISTANCE

Melvyn J Hunt

Phonetic Systems UK Limited
Millbank, Stoke Road, Bishops Cleeve, Cheltenham, England, GL52 8RW
mhunt@phoneticsystems.com

ABSTRACT

Three classes of practical speech recognition dialogue systems are considered, starting with PC-based systems, specifically dictation systems. Although such systems have become very effective, they have not achieved mainstream use. Some reasons for this disappointing outcome are proposed. Speech recognition is now appearing in production cars. It is argued that the two most attractive in-car applications are for navigation systems and for dialing by name. The latter may be more suited to equipment that can be detached from the car and connected to a PC. After considering telephone applications in general, the importance of automated DA (directory assistance – also called directory enquiries or DQ in some countries) is established and its particular challenges are discussed. Among these are the size and dynamic nature of the databases accessed, and the variations produced by callers in naming a commercial/administrative entity whose number they are seeking. The advantages of a bottom-up phonetic speech recognition technique for automated DA are described. It is concluded that the combination of this technique and automatic methods for handling name variation makes automated DA, including access to business listings, a practical proposition.

1. PC-BASED SPEECH RECOGNITION

Applications of speech recognition in which the user interacts directly with a PC and its monitor are dominated by automatic dictation. At first sight, dictation software might not seem to be a dialogue system, but the user engages in a dialogue with a PC in controlling editing, formatting and especially in correcting recognition errors.

The first commercial general-purpose dictation system, *DragonDictate*, appeared just over a decade ago [1]. A remarkable achievement for its time, it was nevertheless by today's standards not easy to use. Users had to learn to pause between each word. They had to read enrolment text in this style for an hour or so. Even then, since the dictation accuracy was not very high, users with no problems with their hands – RSI, arthritis or paralysis – could almost certainly create texts faster by typing.

By contrast, today's dictation software is very effective. Enrolment has been cut down to a few minutes, users no longer need to pause between words, vocabularies are larger, accuracy is much higher, and commands for correcting, formatting and editing have become powerful and flexible.

And yet as a means of creating text automatic dictation has remained a minority activity, almost a curiosity, and there is little evidence of its making much headway to change this situation. It is interesting to ask why. Indeed, for someone such as this author who would find life difficult without automatic dictation – this very text has been created by voice – the question is intriguing.

Over the past two or three years there have been some factors that are not directly related to the usefulness of automatic dictation as a technology. First, much of the progress made in the technology has relied on exploiting increasingly powerful processors and increasingly large amounts of memory. Apart from games, no other software likely to be used on a PC requires the processing power needed to make the latest automatic dictation systems function at their full potential. Although the price of the dictation software fell, if it needed a new computer to use it, it was an expensive proposition. As sales of PCs began to decline, sales of dictation software could not be expected to grow much.

The second indirect factor was that Lernout & Hauspie (L&H), having acquired two key players, Dragon Systems and Kurzweil AI, was found to be in a disastrous financial situation.

These factors cannot explain everything, however. There are also factors directly related to how we view automatic dictation. Dictation competes with the keyboard and mouse. Many people have already put in the effort to learn to type skilfully. Even those who have not done so recognize that fast, accurate typing is a skill that has to be learned and that the effort will be repaid.

By contrast, people feel that they already know how to speak and that no further effort should be needed to use automatic dictation software. The manufacturers have abetted this delusion by choosing names such as *NaturallySpeaking* and with publicity showing users adopting extremely relaxed poses chatting to their PCs as they might to a colleague. In fact, one cannot speak to a dictation system as one typically would to another human being and expect to get accurate recognition. Just as it is necessary to learn how to type effectively, it is necessary to learn how to speak to a dictation system; the difference is that the latter learning process is quicker and easier. Many purchasers of automatic dictation systems, under the understandable impression that no learning was necessary, must have been disappointed and consequently stopped trying to use the product.

In addition to this problem of attitude and education, there remain a couple of practical drawbacks. The first is the need if the environment is even slightly noisy to wear a headset fitted with a gradient microphone and tethered to the PC. One study [2] found this to be the most commonly raised objection to automatic dictation. The second practical problem is the loss of privacy in speaking rather than typing text. This author can personally attest to how inhibiting it feels to have a neighbouring colleague audibly react to a sentence just composed in a text being drafted.

One can fervently hope that the attitude problems among the general public will eventually be overcome and that more convenient microphone arrangements, perhaps adapted beam-forming arrays, will be introduced. These developments are unlikely to occur in the short-term, however. In the near future at least, prospects with specialized applications such as medical dictation may be better.

2. IN-CAR SPEECH RECOGNITION

In-car speech recognition [3] as an installed option for controlling non-critical functions in a car is now being fitted in, for example, the Jaguar *X-Type* and *S-Type*. As an application, it is much more recent than dictation, and it therefore remains to be seen how well users will accept it.

It has the massive advantage over PC-based applications in that it does not have to compete with the keyboard. Indeed, in a car it is desirable to minimize the need for drivers to move their hands from the steering wheel, and speech recognition can help to do this.

On the other hand, a moving car is a noisy environment, and drivers cannot be expected to wear headset-mounted microphones. Moreover, with a system permanently mounted in a car whose driver may change at any time, there is little opportunity for explicit speaker enrolment.

While being able to control the radio or the air-conditioning without taking one's hands from the wheel or one's eyes from the road is both convenient and a safety asset, hands-free dialing will probably have a larger impact. Whether this will improve safety or not is debatable. Certainly, hands-free dialing while driving is safer than using a keypad. However, some reports have argued that it is the telephone conversation itself that poses the most distraction to the driver. Nevertheless, if we are to have hands-free dialing then we need to try to make it work well.

The most convenient way of making the telephone call from a car is to name the person being called rather than reciting a memorized number. As researchers at Nokia have pointed out [4], however, many users have long lists of people that they may wish to call. Training a system in the car for all those names can be tedious. It may be better to make at least part of the speech recognition system detachable from the car and attachable to a PC, where the user's personal telephone directory could be downloaded. Pronunciations would be needed for the names, of course. This topic is discussed in Section 3.2.1. A further advantage of a detachable system would be that the user could carry out some enrolment through the PC.

Arguably the most interesting in-car application of speech recognition is for navigation systems. Such systems offer quite remarkable help in guiding drivers to their destinations in unfamiliar territory. Unfortunately, entering details of the destination through a small keypad or by selecting letters in the alphabet using a cursor is currently extremely awkward. Being able to speak the name of the destination can transform the usability of this application. Although the size of vocabulary needed for this application is much greater than for others in cars, the car need not be moving for speech recognition to be useful. This means that the environment can be much quieter. Also, in a stationary car, the user can be presented with visual feedback, allowing efficient selection from a list.

3. TELEPHONE SPEECH RECOGNITION

3.1 Introduction

Of all the broad classes of applications of speech recognition, applications involving recognition of speech transmitted over telephone links have certainly received the most attention. This is not surprising given that there are vastly more telephones in the world than PCs or cars. Moreover, telephone speech recognition does not have to contend with the keyboard as competition, only the much more limited keypad.

On the other hand, telephones currently offer no visual feedback. This rules out the efficient use of choice lists, which can be exploited by dictation systems and navigation systems to allow the user to select the correct interpretation of what he or she said. Telephone applications typically function with unknown users, ruling out any enrolment process, and the interactions are too brief to allow much adaptation to the caller's voice. Finally, and most obviously, telephone links impose bandwidth limitations, and can add noise and distortions. Although the signal quality with fixed-line telephone links is getting better, the continually increasing popularity of cellphones imposes evident new challenges for ASR. The effect of coders used in cellphones appears reasonably manageable. The problem, rather, is that cellphones are frequently used in much noisier environments than fixed-line telephones, and the interaction between the background noise and the coder can be particularly challenging, as can the effects of signal fading with a caller on the move. It remains to be seen whether distributed speech recognition approaches such as those being explored by the ETSI-Aurora project [5], in which the acoustic analysis for the speech recognizer is carried out locally in the telephone before transmission, will take off.

Early, simple applications of telephone speech recognition in interactive voice response (IVR) applications typically required the user to speak digits together with a few words such as yes and no. They consequently competed directly with the keypad and were not so compelling. Now we have applications ranging from automated personal assistants to voice browsers, attempting to approximate Internet browsers. There is also obvious potential for intelligence and security applications, though any deployment of such systems is unlikely to be made public.

The rest of this section focuses on an application of telephone speech recognition, which is particularly challenging but also of great commercial importance, namely automated DA.

3.2 Automated Directory Assistance

The provision of directory assistance services is a huge activity. There are said to be around 6 billion DA calls per year in the US [6]. Currently, just over one in six of those calls is from a cellphone, but while the number of calls from fixed-line telephones is static, as office-based callers increasingly use the Internet to find phone numbers, calls from cellphones are growing at more than 20% per year. In the US a DA operator's position (occupied by different operators over the day and over the week) costs around $150,000 per year.

In Europe, there are estimated to be around 2.8 billion DA calls per year [7], with a similar pattern to the US in the proportion of calls from fixed line and mobile telephones. Former monopoly telephone service providers are generally required to maintain a DA service, but under EU regulations they are not allowed to cross-subsidize the service, and it must therefore pay its way through fees charged [8]. The annual revenue from the fees charged for DA calls, which range from about 35¢ to around $2, are estimated to be $1.75 billion in Western Europe alone. In addition, most telephone service providers in both the US and Europe gain additional revenue by charging a further fee for connecting the caller directly to the number requested (so-called "call completion".)

In the UK, which has the largest DA activity in Europe with around 800 million calls, the national regulator, Oftel, has just announced that the BT 192 number will be withdrawn and is to be replaced by a set of new five-digit numbers (118xy), allocated to perhaps 10 or 20 competing DA service providers.

Developments such as these and the sheer size of the activity make it very attractive to reduce costs by automating or at least partially automating the service. In parts of the US, callers already have effectively no human contact with the operator: recorded prompts ask them questions, the callers' responses are recorded and played to an operator, and the telephone number found is played automatically to the callers. The obvious next step is to automate the process entirely by having a machine recognize what the callers say.

Automated DA is different from other applications of speech recognition in that speech is not functioning as an alternative input mode: there is no practicable keyboard or keypad alternative to compete with it. Rather, speech recognition is replacing the human operator, and if it can be done well enough callers will be unaware of any change from the current semi-automated services.

Automated DA is unlikely to replace human operators completely because there will always be more complex inquiries where human skills are necessary. However, we may hope that it will replace the basic, increasingly depersonalized, operator's job, which is repetitive and unrelenting – operators in the US are required to dispose of callers within 20-25 seconds on average.

3.2.1 The Challenges of Automated DA

DA calls can be split into requests for residential numbers, identified by the name and the address of the subscriber, and requests for so-called business numbers. (Business numbers are sometimes called "Yellow Pages" numbers, but this seems a misnomer since Yellow Pages are classified by the type of business, whereas DA services normally function only on the name of the business.) Business numbers encompass government services, schools, hospitals, *etc*, professionals such as doctors, dentists, lawyers, *etc*, as well as the more obvious commercial listings, such as restaurants, shops, offices, factories, *etc*. Four out of five DA calls in the US are requests for business numbers, and the proportion is similar in the UK.

DA calls are similar in many ways to calls to a large autoattendant covering a complete enterprise (such as, for example, those installed by Phonetic Systems at Motorola, the Bank of New York, Atlantic Records, *etc.,* accessing directly up to 200,000 named employees). However, callers using a public DA service are drawn from a much wider population, including the young and the very old, compared with those using an enterprise system. Furthermore, callers to a DA service pay a fee for the call and have a right to expect a reliable service. Indeed, their interests are often protected by a regulatory body ensuring that the service maintains an acceptable standard.

Callers frequently cannot spell the name of the person or business they are seeking. They may make mistakes in other information that they provide. In particular, they may misidentify the location, ascribing it to a neighbouring town or to the metropolitan city rather than the suburban town where it is in fact located. This means that an effective system cannot expect to take in each piece of information and move on to the next. For example, if a caller has confirmed that she believes that a number is located in the London suburb of Edgeware, it is not a good strategy to look only at Edgeware numbers, since the number may well turn out to be located in the neighbouring suburb of Stanmore.

Predicting how callers will pronounce proper names is a major challenge. Even humans have problems with this task, particularly with names coming from another linguistic community. It is not enough to know how that community would have pronounced the name, because it will be adapted to the local language to an extent that is often difficult to predict. For an automatic speech recognizer, and for the text-to-speech system that must speak it back, this problem is particularly difficult. If not specially corrected, an automatic system will assume that a name of Italian origin in the US such as Marchese should be pronounced /m AH r h iy z/ rather than the usual American pronunciation, closer to the Italian, namely, /m ah r k EY z ih/. In principle, such problems could be handled by having very large dictionaries. Deciding the pronunciation of business names, however, can pose an additional problem because they may include a unique made-up name such as *Kleen-EZ* (clean easy), for which the deduction of the pronunciation needs a different strategy.

Residential numbers pose problems due to the sheer size of the database to be searched (more than 150 million residential listings in the US). There are also problems with first names that may be replaced by standard variants: Bob/Bobby/Rob/Robert, Betty/Beth/Liz/Lizzie/Elizabeth. Worse still, the directory listing may use an initial, while the caller may cite the subscriber's full first name.

We have seen that it will be more useful to be able to handle business number enquiries than those for residential listings because there are many more of them. At first sight, business name DA automation may appear easier because there are fewer names to choose between (about half a million on average for a US state that might have perhaps 5 million residential listings). However, business names pose a large and fascinating problem, namely coping with the variations produced by callers on the name as it appears in the listing.

Let us look at some examples. The former well-known speech recognition company Dragon Systems, Inc. may be called Dragon, Dragon Systems, Dragon Systems Incorporated, Dragon Systems "ink", or, bizarrely, by the name of one of its products, DragonDictate. The major British telephone service provider is officially known as British Telecommunications PLC. It also uses BT as an acceptable abbreviation. However, it is probably most commonly referred to as British Telecom, a completely unofficial form. Synonyms appear in caller's enquiries. For example, the terms "hospital," "clinic" and "medical center" get freely mixed. The location of a business may appear in its official listing name, for example, the Sheraton Needham. Callers are likely to behave differently and add the location information when it is not in the listing name of leave it off when it is. Words frequently get reordered by callers relative to the listing name. Is it, for example, The Hotel Belvedere or The Belvedere Hotel; The University of Oxford or Oxford University?

Perhaps the most challenging business name problems are associated with large organisations with many departments: a level of government, a university, or a multinational corporation. It is in general impossible for the caller to know how the organisation has listed its departments: if, for example, the caller is looking for employment with a large organisation, should he ask for The Personnel Department, Recruitment, Hiring, Human Resources, HR...?

For the businesses that are requested most frequently, it is both practicable and desirable to generate common variations on the listing names manually, ideally by observing directly what real callers say. This is of course not practicable for the vast majority of listings. Researchers in Italy [9] have reported work on automatically detecting name variations used by callers for specific listings. Our approach at Phonetic Systems is somewhat different. By studying tens of thousands of real calls we are developing models of how variations can be statistically predicted and handled in the recognition process.

Despite the extraordinary complexity in the way that callers generate name variations, we are finding that the problem is tractable. When the automatic techniques for handling business name variations are combined with the phonetic decoding process described in the next section, our tests indicate that a useful level of recognition of business names as spoken by real callers can be achieved

3.2.2 Phonetic Decoding for Automated DA

Back in the 1970s it was widely assumed that the only way to achieve large-vocabulary speech recognition would be through a bottom-up process (Fig. 1). Speech would first be segmented into phonemes, the phonemes would then be identified, and words would finally be recognized from the phoneme sequences. One early publication in an august journal described a large-vocabulary speech recognizer, with detailed descriptions of most processes but with a magic box in the middle called the "phoneme decoder" out of which came a sequence of phonemes. I believed this to be seriously misguided. There are no clear boundaries between most phonemes, the acoustic realisation of most phonemes depend heavily on their context, and there is simply not enough acoustic information in the part of an utterance corresponding to one phoneme to allow it to be accurately identified, even by a human listener let alone by a machine.

The *Harpy* system at CMU [10] was an early example of a successful alternative approach to large-vocabulary (continuous) speech recognition. The phoneme sequences occurring in the words in the vocabulary and the allowed word sequences in the grammar were compiled into a single network used directly by the speech recognizer (Fig. 2). In the 20 years that have followed, this approach has been dominant in successful speech recognition systems, resulting in remarkable achievements as exemplified in some of the systems evaluated in the DARPA trials [11] and in commercial products such as *Dragon NaturallySpeaking*™. For most of that time I believed that it was the only practical approach to large-vocabulary speech recognition. I now know that I was wrong.

For some applications it now seems clear that as a result of various advances a bottom-up approach in which phoneme sequences are generated without initially applying word-based constraints has become not only practicable but is in fact preferable to the conventional top-down approach. Over the decades our ability to generate acoustic features that map to phoneme identity has improved significantly [12,13]. We have learned to mitigate context-sensitive variation in phoneme realisations by using context-sensitive phoneme models [14] and by representing their spectra by Gaussian mixtures rather than by a single simple parametric distribution [15]. We can mitigate the problem of uncertainty in phoneme boundaries by generating not a single segmentation into phonemes, but rather a phoneme lattice [16]. Finally, we can use our knowledge of the probability of confusions between similar phonemes (s/f, m/n, for example) to mitigate the inevitable acoustic ambiguity in the speech signal.

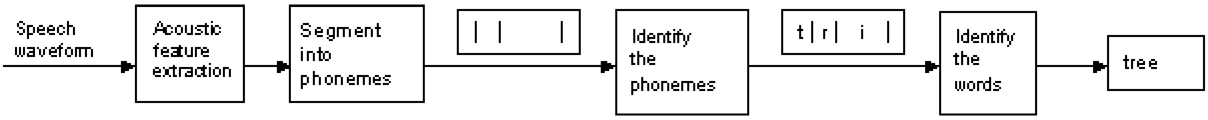

Figure 1. An early naïve view of how a large-vocabulary speech recognizer might work. The speech waveform is first split into phoneme units without reference to the identities of the phonemes. The phoneme identity of each unit is then determined. Finally, the phoneme sequence is matched against the known sequences for words in the vocabulary, possibly taking phoneme confusion probabilities into account.

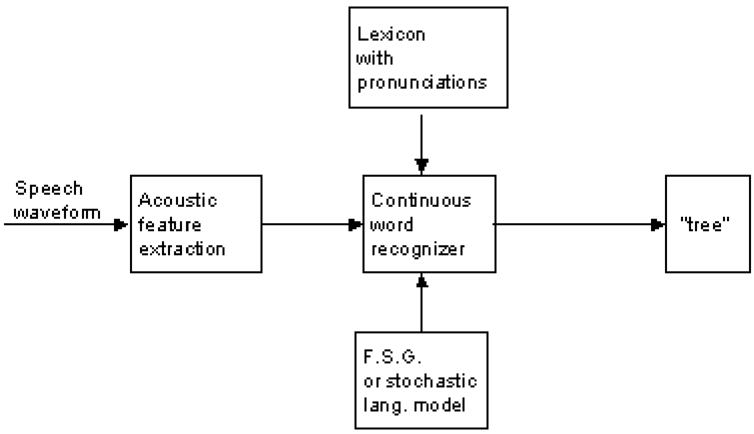

Figure 2. Structure of a typical modern speech recognition system. Knowledge of all words that can be recognized, and any sequence constraints, is built into the recognizer itself. The output may be a single word or word sequence. In other cases it may be a list of so-called N-best interpretations of the input that can be combined with results from other inputs to decide on a single best interpretation.

A phoneme decoder with the features just described (Fig. 3) turns out to be well suited to accessing very large, dynamic databases, specifically DA databases. Such a decoder generates a phoneme lattice from each utterance that the caller makes. The database contains one or more phonetic transcriptions of each lexical item in it. Using knowledge of phoneme confusion probabilities, the phoneme lattice can be used to compute the probability estimate that the caller spoke any item in the database. Note that the decoder is completely decoupled from the database and its computational load is independent of the size of the database.

The alternative, more conventional, approach would be to try to compile the complete database (or rather the part containing all the last names, *etc*) directly into the speech recognizer. Leaving aside the question of the feasibility of this integration and the accommodation of continual updates, such a system will only generate a finite number (say, 10 to 50) of possible interpretations with associated probabilities. As we have noted, effective DA automation requires the integration of information from multiple utterances: in the case of residential listings, this might be the first name, last name, city, perhaps some spelling, perhaps some address information, perhaps some items repeated. It is conceivable that, say, the correct first name and, say, the correct city are both the 51st best interpretation or worse, yet when combined with the other information obtained, they contribute to the choice of the correct listing. A word-based recognizer that cut off after 50 interpretations would not be able to use them.

The phonetic approach may also have another advantage for automated DA. As we have seen, it is practically impossible to predict how callers will choose to pronounce all names in a database. Attempting to encode all possible pronunciation variations in a word-based speech recognizer would lead to a combinatorial explosion, and in any case experience has been that including multiple pronunciations is of little value or is even harmful unless they are augmented with prior probabilities [18]. The phonetic approach, however, appears to be more tolerant of unanticipated pronunciations. In a traditional word-based recognizer using beam pruning, there is a distinct possibility that a phoneme mismatch, particularly a stressed vowel, will cause

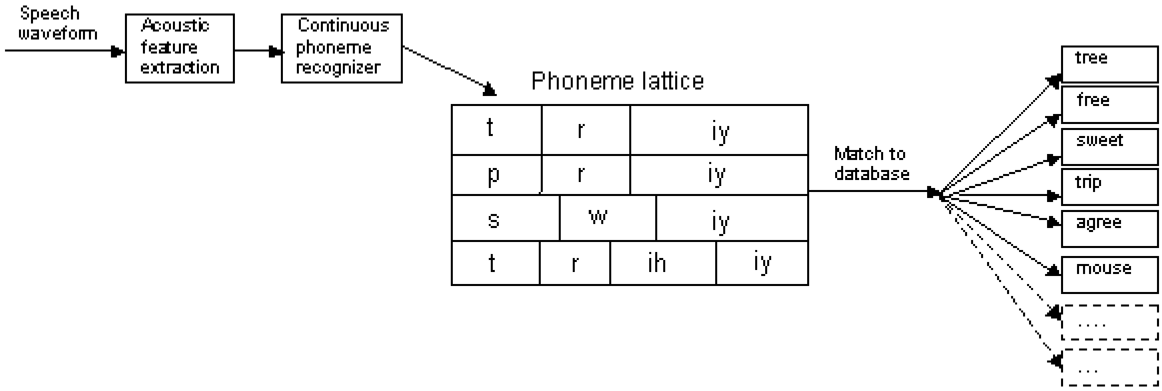

Figure 3. A phonetic recognition system suitable for accessing large databases. The recognizer itself has no knowledge of the words in the database. It generates a phoneme lattice, which can be matched, using knowledge of phoneme confusion, deletion and insertion probabilities, against every item in the database. The match scores can be combined with information from other inputs in a dialogue designed to determine the speaker's intention. The phonetic match process is tolerant of differences between the speaker's pronunciation and the expected pronunciations in the database.

the correct word to be dropped from the beam. The phonetic approach is not subject to this danger. Moreover, it is possible to score hypotheses with the phonetic approach in a way that does not automatically weight the contribution of any phoneme by its length, as is the case in word-based recognition. In this way, the influence of long, stressed vowels, which are often a source of pronunciation variation, can be kept within bounds.

The ability of the phonetic approach to provide a probability estimate for any interpretation of each utterance and then combine those estimates across utterances makes it well suited to sophisticated dialogue structures. The system can decide at each point whether to ask for new information, ask for confirmation, ask for information to be repeated, announce a number or decide to pass the call to a human operator.

It is true that if the expected pronunciations are correct, and if the word-based recognizer could be made to integrate the complete vocabulary, then for *individual* utterance recognition it may outperform the phonetic approach just described. However, the phonetic approach does not exclude the use of computationally efficient rescoring of the top candidates using word models as in the conventional approach. In this case, the phonetic recognizer acts as a kind of rapid match process. Note, though, that with the phonetic recognizer the system can first integrate information from candidates at arbitrary depths and decide in the light of that integration which hypotheses should be rescored.

For automated DA, the phonetic approach consequently seems to have no inherent disadvantages over the traditional word-based approach and several crucial advantages.

4. CONCLUSIONS

Of the three application areas of speech recognition that have been considered PC-based dictation systems are the most mature and represent a remarkable technical achievement. They have, nevertheless, failed so far to gain really widespread use because they are competing directly with the keyboard and because of

unrealistic public expectations. Speech recognition in cars for hands-free dialing and command-and-control applications has already appeared, but applications in navigation systems and with removable mobile phones look more promising. Automated directory is a commercially important application, especially for business names, and there is no practical keyboard or keypad alternative. In this application an unconventional phonetic approach offers major advantages in searching large, dynamic databases within a dialogue where several questions may be asked and where there is uncertainty about how proper names will be pronounced. With the phonetic approach to speech recognition and the automatic handling of variations in business names currently being developed, the prospects for effective automatic DA look good.

5. REFERENCES

1. J.M. Baker, "DragonDictate™-30K: Natural Language Speech Recognition with 30,000 Words," *Proc. ESCA, European Conference on Speech Communication and Technology*, *Eurospeech 89*, Vol. 2, Genoa, Italy, September 1989.

2. Hunt M.J., "Practical Large-Vocabulary Speech Recognition in a Multilingual Environment *Speech Communication,* Vol. 23, No. 4, Special Issue on Non-Telephone Applications, eds M.J. Hunt and F. Néel, December, pp. 297-306, 1997.

3. M.J. Hunt, "Some Experience in In-Car Speech Recognition", *Proc. IEEE/Nokia Workshop on Robust Methods for Speech Recognition in Adverse Conditions,* May 25-26, 1999, Tampere, Finland, pp. 25-32.

4. K. Laurila and P. Haavisto, "Name Dialing – How Useful Is It?", *Proc. IEEE Int. Conf., Acoustics, Speech & Signal Processing, ICASSP-2000*, Istanbul, Vol. VI, pp. 3731-3734.

5. Aurora project website: http://www.etsi.org/technicalactiv/dsr/dsr.htm

6. Information provided by The Kelsey Group.

7. European Directory Assistance Markets, The Pelorus Group, July 2001.

8. Smada project website: http://smada.research.kpn.com

9. C. Popovici, M. Andorno, P. Laface, L. Fissore, M. Nigra and C. Vair, "Learning of User Formulations for Business Listings in Automatic Directory Assistance", *Proc. EuroSpeech 2001*, Aalborg, Denmark, pp. 2325-2328.

10. B. Lowerre and R. Reddy, "The Harpy Speech Understanding System" in *Trends in Speech Recognition*, W.A. Lea, (Ed.), Prentice-Hall, Englewood Cliffs, pp. 340-360, 1980.

11. D. S. Pallett, J. G. Fiscus, A. Martin, and M, A. Przybocki, "1997 Broadcast News Benchmark Test Results: English and Non-English", *Proc. Broadcast News Transcription and Understanding Workshop*, February, 1998, Lansdowne, Virginia, pp. 5-9.

12. N. Morgan, "Temporal Signal Processing for ASR", *Proc. IEEE International Workshop on Automatic Speech Recognition and Understanding (ASRU)*, Keystone Resort, Colorado, December 12-15, 1999.

13. M.J. Hunt, "Spectral Signal Processing for ASR", *Proc. IEEE International Workshop on Automatic Speech Recognition and Understanding (ASRU)*, Keystone Resort, Colorado, December 12-15, 1999.

14. K.F. Lee, *Automatic Speech Recognition — The Development of the SPHINX System*, Kluwer Academic Publishers, Boston, 1989.

15. B.H. Juang and L.R. Rabiner, "Mixture autoregressive hidden Markov models for speech signals," *IEEE Trans. Acoustics, Speech, Signal Proc.,* ASSP-33 (6): pp.1404-1413, December 1985.

16. J. Wolf and W. Woods, "The HWIM Speech Understanding System" in *Trends in speech recognition*, W. A. Lea (Ed.), Prentice-Hall, Englewood Cliffs, pp. 316-339, 1980.

17. B. Peskin, M. Newman, D. McAllaster, V Nagesha, H. Richards, S. Wegmann, M. Hunt and L. Gillick, "Improvements in Recognition of Conversational Telephone Speech," *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, *ICASSP-99*, Phoenix, Arizona, Vol. 1, pp. 53-56.