Abstract
Large-vocabulary continuous-speech recognition (CSR) technology is at work. As an application of the technology, we will describe a dictation system (DS). Input to the system is unrestricted spontaneous speech. No adaptation, no special skills are required to use the system. The DS transforms continuous speech into written text. It is essential in this application that the user is free to speak as he or she usually does and should be free to use his or her own wording and formulation. This implies speech recognition for large and open vocabularies, free syntax, continuous speech. The aim of the paper is an attempt to determine what is feasible with today's technology and what will be feasible in the near future. The problems addressed are: what are the limits of today's technology, what is needed to make the next step, i.e. going towards real industrialization of CSR technology.

Keywords: Continuous-speech recognition; free syntax; dictation system; vocabulary selection; on-line adaptation; domain.

1. Introduction
Continuous-speech recognition (CSR) technology is an important subset of language technology and has started rendering new products feasible that will revolutionize not only man–machine interfaces but also information access and processing. The aim of the paper is an attempt to determine how far we have gone today (i.e. what is feasible with today's technology) and how far we can go tomorrow (i.e. what will be feasible in the nearest future). This will lead to the questions: what are the limits of today's technology? What is needed to make the next step (i.e. going towards real industrialization of CSR technology)?

The paper tries to undertake this study with the help of an application that can be considered representative of the state-of-the-art of the technology:
large-vocabulary dictation for which Philips Dictation Systems has offered, since June 1994, a product (SP6000) for the German language. But first, we discuss what is a large vocabulary.

Next, we describe what are the impacts on a user of this new technology. An analysis follows that attempt to answer the questions: where are we, what is actually feasible? What are the limits of the actual technology and what do they imply for a product? What resources have to be tapped, which actions have to be initiated to be able to advance further?

2. What is a large vocabulary?

The word 'large' is undefined and, in general, strongly depends on what we are talking about. For speech recognition the notion of 'largeness' has evolved over time. In 1989, some American research groups [1] as well as Philips Research [2] were proudly demonstrating 'large vocabulary' continuous-speech recognition systems of about 1000 words that allowed one to make database queries following a very formal syntax. In 1995, the same groups are still demonstrating 'large vocabulary' recognition systems, but now with a vocabulary of 65000 words and an unconstrained syntax.

Since 1994, the notion of 'largeness' has been very much related to 'infinity' [3]. A large vocabulary is now an unlimited or open vocabulary, introducing the notion of out-of-vocabulary (OOV) words. An OOV is a word that has been pronounced by a user and which does not belong to the vocabulary pre-selected by the system. The OOV rate tells us how well the recognition vocabulary covers an application. The problem is that, due to the nature of speech, an OOV rate of 0% will never be reached. Typically an OOV rate of about 0.5% is striven for in an application, leading to a vocabulary of 65000 words for North American Business (NAB) news dictation and 8000 words for English radiology report dictation. Moreover, there are differences between languages. In German, a vocabulary of more than 100 000 words is needed to have a good coverage of newspaper article dictation, and 25 000 words are necessary for radiology reporting.

The size of a large vocabulary depends then very much on both the language and the type of application it should cover: it ranges between 5000 and more than 100 000 words, but it has the property of covering as much as possible, not necessarily all of the real-world application's vocabulary. On the contrary, a small-vocabulary recognition system is used to issue commands as is done e.g. in vocal dialling for a car telephone. The vocabulary should include all command words, that is, a 100% coverage is a must. For an inquiry system, like the train time-table information system developed by Philips, an
Large-vocabulary recognition

interpretation of the recognition output is made because some words are only noise and need not be recognized. Therefore, a complete vocabulary coverage of all the words used to make a query is not necessary. Here, apart from city names, a coverage of about 90% should be sufficient.

To resume, a large-vocabulary recognition system is a system that has to recognize all words within a domain of application. Therefore it needs an unlimited or open vocabulary, i.e. a vocabulary that covers the domain of application by as much as by 99%. In the next section, we consider an example of such a system, the SP6000, the continuous-speech recognition system from Philips Dictation Systems.

3. A dictation application

A system that generates a written document from a voice input is the first product everybody has in mind when thinking of a possible use of the CSR technology. Managers or doctors have already partly had this dream fulfilled of having very quickly on paper what they have just said when dictating a letter or a report to a secretary. The intention behind using an automatic transcription system or dictation system (DS) for these dictation professionals is to improve transcription efficiency and to produce a final document quicker with fewer errors.

![Diagram of traditional dictation scheme](image)

Fig. 1. The traditional dictation scheme in a hospital.
What are the different processes involved in the creation of a document by voice? Fig. 1 shows how it is traditionally organized in a hospital. The doctor dictates using a dictation device in his office. His voice is recorded on a magnetic tape (generally a cassette) that travels then to the typist’s room. When the cassette has been transcribed, a written text will make the same journey back to the doctor or the patient’s file. A fully digital system however, where the dictation is routed by a network directly to the typist’s PC, certainly helps to solve problems like, e.g., cassettes getting lost on their journey to the typist’s room, but it does not reduce the workload of the typist.

It is not very difficult to imagine extending the digital system to a new system (Fig. 2) where a continuous-speech recognizer (that automatically transcribes the dictation) is added to the network. Note that the typist has not disappeared. The document produced by the recognizer will never be error-free and is probably not suitably formatted. The typist now has to verify and from time to time correct the output of the recognizer as well as apply some refinements to the final document.

What are the technology requirements of such a product? Certainly the habits of the person dictating cannot be changed as the overall document creation process should increase in efficiency. This means that:

- The dictation should be made possible in a fluent way, at a normal speed without having to observe artificial pauses after each word.
• Hesitations should be allowed.
• Recording might occur in noisy environments.
• There should be no restrictions on the vocabulary employed.
• The syntax should be free, i.e. technical expressions that are grammatically incorrect have to be recognized.
• No use should be made of any keyboard by the person dictating.
• Formatting information given by the person dictating should be taken into account (like 'new line', 'paragraph', 'period').
• The final document should have at least the same quality, if not a higher one, than without DS.
• The correction of the recognized dictation should be made with a comfortable speech-synchronous editor.
• The system should run on inexpensive hardware, such as a PC.
• Time response (i.e. the process time for transcribing the dictation) should be near real-time.
• Working conditions should increase in quality, especially for the typist who should have a comfortable environment to correct the document transcribed by the recognition system.

Is this realistic? Under certain conditions (we will come back to this point in section 6), yes! A radiologist dictating a medical report in front of his X-ray pictures is satisfying these conditions. In June 1994 Philips Dictation Systems launched, for German speaking radiologists, a DS called Speech Processor 6000 (SP-6000). This is the first and until now (February 1995) the only real product that is on the market able to recognize large-vocabulary continuous-speech. The system has already been installed in several hospitals in Germany and Austria. The system is speaker-dependent (requiring about 40 minutes training data), recognizes continuous speech, it accepts a vocabulary of up to 28,000 words and the syntax is given by a domain-specific bigram language model. Some automatic retraining procedures allow the system to reach optimal performance after several hours of dictation have been produced by the speaker. The system runs on a 486 PC with a 66 MHz clock rate and two additional boards, the first one carrying a signal processor that performs the acoustic analysis, the second with a dedicated co-processor that contains a gate array IC to accelerate the recognition process. The PC can be embedded in an already existing dictation environment as in Fig. 2. As already mentioned, this product is now available for the German language and will soon be on the American market.

The performance of this system is fairly satisfying. It transcribes the
dictation with a real-time factor\(^1\) that varies from speaker to speaker between 1.0 and 1.5 with a word error rate less than 10%. This means that the typist receives on his/her screen a text that is more than 90% correct. The overall system performance is even more interesting: a speed up of up to 40% has been observed in document creation. In addition, the documents now contain fewer errors.

4. Impact on users

Now, the question arises of how users will react to this new technology, what are the consequences to be observed. But who is the user? Is it the secretary or the doctor? Will the doctor himself want to correct the dictation? Does it make sense for the doctor? Will technology increase the quality of the created documents, or will documents lose in quality? There are many questions like these for which we have answers, but because of a lack of experience (the technology is just emerging on the market) these answers are tentative. Moreover, users very often have unpredictable attitudes towards new products. So, what will follow in this section is our view, from Philips Research, on user-impact aspects and for which we have prepared our reactions.

The DS procedure according to Fig. 2 allows us to preserve the document generation structure of the department in which it is installed. This means the doctor dictates at the same place as usual; the typist or secretary keeps his/her position. The old structure is kept, only the way work is done is somewhat different.

From the point of view of the doctor, everything is perfect: he dictates as usual, with the same microphone having the same old features in backwarding, forwarding and listening facilities. In addition he is pleased to work with a high-tech product. Moreover, he receives the written document more rapidly.

The typist has to change habits. The work does not consist any more in menial typing but in verifying that the transcription is correct. It implies another kind of concentration that puts the typist's qualification at a higher level with increased responsibility.

Philips marketing people are expecting a global efficiency improvement in document creation by up to 40% with a higher quality of documents having fewer errors. Moreover, the system is available at weekends, nights and during holidays: there is also a gain in flexibility.

\(^1\) This is the quotient of processing time over input time; i.e., a real-time factor of two signifies that the processing of a seven-second-long input needs fourteen seconds. Ideally, for speech recognition, a real-time factor of less than one should be striven for.
Large-vocabulary recognition

Everything seems perfect and the sales should grow exponentially. But large-vocabulary recognition systems also have weak points that we should be aware of, as they show us that we are just beginning to tackle and therefore to understand industrializing speech recognition.

5. Shortcomings of large-vocabulary recognition systems

With regard to continuous-speech recognition technology, Philips Research has been shown to belong to the world’s leading laboratories [3]. Unfortunately, the technology itself does not make people buy a dictation system. It is a new challenge to make a product based on this technology. A large-vocabulary system is open to the world. Its vocabulary will evolve over time, new words will be added to it, others will never be used again. Further, within a domain like radiology, variations in terminology are observed from hospital to hospital. The user’s grammar is also very flexible, not necessarily following rules. Modelling has to be robust against these natural variations of speech, between and within users. Moreover, such a system will never produce an error-free output. A good environment including easy-to-use facilities has to be supplied with the recognizer in order to help the user to maintain his system. The system should also be able to adapt to the user. In this section we will discuss the points that make a DS vulnerable while opening it to the real world.

5.1. Vocabulary selection/coverage

The problem begins with the initial selection of the vocabulary. A short example shows the importance of vocabulary coverage for an application. During the past five years (from 1987 until 1993), a tremendous amount of financial articles from six different North American newspapers has been collected to make the so-called NAB news corpus. The corpus totals 240 million words, 446,000 of them are different. These 446,000 words build what we could call the background vocabulary.

Do we need a vocabulary as large as this to make speech recognition feasible for NAB news dictation? Actually, a subset of this vocabulary should be sufficient, and an optimal coverage has to be found. To evaluate the coverage of such a lexicon is very simple. We need only to select articles from a new period

\[\text{See also section on benchmarking in [4].}\]

\[\text{Actually, 120 million words have been acquired between 1987 and 1993 from the Wall Street Journal (17 million words per year), 110 million words between 1988 and 1990 from the Agency Press (36 million words per year) and 10 million words from the San José Mercury in 1991.}\]
(recent ones, from 1995) and count those words that appear in these never-seen articles and that are in the vocabulary. With a vocabulary containing the 20,000 most frequent words seen in the period 1989-1994, a coverage of 97.5% of the articles is assured. This means that 2.5% of the words of a new article are not among the 20,000 most frequent words extracted from the five years’ experience with the domain. Is this coverage high enough? It has been observed that for each word that has been dictated and that is not in the vocabulary, two errors or two mis-recognitions will be produced. Hence, 2.5% words of a dictation not in the vocabulary gives rise to a base word error rate of 5%\(^4\). In addition to this there is the standard word error rate (between 5% and 10%), which makes a total word error rate of 10% to 15%, which is too high for a product. A solution to this is to increase the vocabulary size. For NAB, the optimal figure seems to be 65,000 words, which leads to a coverage of 99.5%, i.e. a base word error rate of 1%. A 100% coverage is never possible (just think of proper names). For example, the total 446,000 NAB words (covering to 100% five years of NAB text) will cover 99.9% of recent articles. Moreover, speech belongs to those natural events that evolve over time. The coverage of a pre-selected vocabulary will decrease over time, certainly not much but enough to irritate a journalist who will repeatedly have to correct the name Bill Clinton while dictating an article on the White House, because the vocabulary had been set up at a time where Clinton was a non-name and did not appear often enough in articles to be taken in the base vocabulary. This leads us to discuss a feature that every large-vocabulary system should have one way or another, namely having the possibility of adding new words to the vocabulary.

5.2. Inserting a New Word

There are two types of problem encountered with the insertion of a new word. First, how should it be done; secondly, how can we ensure that the word will be recognized in the future? To the first question, how should the addition of a new word be monitored, the answer is: the simpler, the better. For example, the user could record the new word and type its spelling with the aid of some tool, and the work should be done.

To the second question, how can we ensure that the system effectively learns the new word, the answer is not trivial. The word has to be modelled acoustically but also has to fit in the grammar. We recall that, as it is impossible

\(^4\) The base word error rate is, in other words, the minimum word error rate the user is certain to have while making a dictation. These errors are related to the words not in the vocabulary of the system, which obviously will never be recognized.
Large-vocabulary recognition

for the system to learn the acoustics of all separate words that might be pronounced at some time, modelling is made on so-called phonemes, the atom of speech [4]. There are around 40 to 50 phonemes for each language, which seems to be a reasonable number. With this set, all possible words of a language can be produced. With phoneme models, the modelling of a word, even of a new word, is done at the symbolic level. A word is a sequence of phonemes, which means that its modelling is done in concatenating phoneme models or, in other words, in the setting of a sequence of phoneme models.

As all phonemes that can ever be pronounced are correctly modelled, the insertion of a new word in the vocabulary implies the handling of phoneme transcriptions which can automatically be done with a tool that, given the orthography of a word, derives the phoneme sequence [5]. This works perfectly well if the words to transcribe are 'normal' words, for which the rules apply. But a proper name, let alone a foreign name, is often 'mis-pronounced' in the sense of the rules. Here the recording of a new word (done while acknowledging the system with the new word) plays a role and some fine-tuning is necessary to cleverly combine knowledge from the transcription tool with the analysis of the new spoken item.

Concerning the grammar or language model, the word receives a minimum probability to start with. As the dictation system is steadily analysing the output of the recognizer, the probability of using this new word will increase over time if the word is frequently used. We can say that the system is learning, learning about new words, their acoustics and also their grammatical meaning. Can the system learn or improve on old words? We call this kind of learning on-line adaptation, which is a very interesting feature enabling flexible use of the system for new users.

5.3. On-line adaptation

We have shown at Philips Research that about ten hours [4] of speech from a given speaker are necessary to train our acoustic models in order for them to be robust. Furthermore, to reach good performance, our language models should have been evaluated on application-dependent texts that total 40 million words, making the equivalent of 150 books of 500 pages each. This means that the making of a speech-recognition product has to cope with a dilemma: being as general as possible to enable as many users as possible to buy the system and, on the other side, being as specific as possible to reach the best performance.

As always for speech recognition, there are two dimensions to consider: speech and text: speech to evaluate our acoustic-phonetic models, text to build a language model, the grammar of the application [4]. The more specific to the
user and the task that these models are, the better the performance of the system. A partial solution to the dilemma is to make the system learn from the user, which we call on-line adaptation. If the texts used to build the grammar are application dependent and — what is interesting — speaker independent for a given application, speech varies very much from speaker to speaker and a new user cannot be asked to record ten hours of speech to train his recognition system before he can use it.

Acoustic-phonetic on-line adaptation is first intended to allow a new speaker, starting with speaker-independent references, to use a recognition system directly without having to train it and with rapidly increasing performance. The speaker-independent acoustic models will be slowly adapted to the speaker's specific voice, building references that are very near to what the system would have learned had it had ten hours of a speaker's voice to train on. Secondly, during the whole life of the system, it will adapt to slight changes in the way users speak, for example the speaker having caught a cold.

Adaptation can also be done on the language model, even though the grammar is speaker-independent for a given application. But some small effects have to be taken into account such as the appearance of new proper names or the peculiar use of grammar by the user. The dynamics of the speaker's vocabulary and grammar have to be evaluated and hence the grammar will adapt to the speaker [6].

5.4. Going across applications

We can adapt speaker-independent acoustic models to the user's specific voice; we can also adapt or improve the user's grammar while incorporating new words. Can we make a bigger step, adapting the system across applications, from one application to another, from newspaper dictation to medical reporting? This is a very difficult question and research work has to be done in this direction.

The very first problem is vocabulary coverage. It gives an indication of how well an application is covered. This coverage problem is already difficult to solve for one given application. Now, taking two very different applications, cardiology and NAB news dictation for example, the problem is more acute. For English, a typical vocabulary size for cardiology is 8000 words, for NAB news 20000 words or, better, 65000 words, Let us first have a look at the lexicon coverage, which is a static coverage; only 58% (resp. 74%) of the 8000 cardiology lexicon is in the 20000 (resp. 65000) word NAB lexicon.

---

3 Work on this topic has still not been published.
A more interesting question for us is how much of what is effectively dictated in cardiology is covered by the NAB lexicon. This is what we call a frequency-weighted coverage that tells us the out-of-vocabulary (OOV) rate. The figures are better, about 80% (resp. less than 90%) of each dictation is covered by the 20,000 (resp. 65,000) NAB lexicon. We have then an OOV rate of 20% (resp. 10%) while using NAB news vocabulary for cardiology. This confirms that a certain core vocabulary (2000 to 3000 words) is used by everybody and covers a large amount of all possible corpora.

This image is very similar when going from NAB to radiology, if not worse. If we shift from radiology to cardiology it is not so critical. But what about shifting to general business letters or to private letters? The background lexicon that reasonably covers such general applications has to be huge.

The reader may remark that to extend a given vocabulary should be rather simple if you list all words in a good dictionary, for example the Oxford English Dictionary for English. This is true. But, in practice, we have to extend our vocabulary from representative real-life dictations collected on the chosen application. Not to do so would make our life more difficult. We have to model word dependencies that go into our grammar model. And the dynamic lexicon coverage, besides the perplexity, is a good indicator that tells us whether we have enough text material first to list a vocabulary with reasonable coverage, but much more to evaluate grammar correctly.

This does not mean we have to collect a tremendous amount of data for each new application. There are actually methods of profiting from the knowledge acquired on one application before switching to another one. Here, adaptation plays an important role.

5.5. Going across languages

Being multilingual is a very much desired property for speech recognition, but not easy to implement. Each language has its own peculiarities. From one language to the other, not only the words but also the phonemes are different (the French nasals are not present in English, nor in German); speaking style evolves. Moreover, language-specific features need a special technology answer. At the acoustic-phonetic level they are named co-articulation in English (the sound of a phoneme is very much influenced by its neighbour phonemes), 'liaisons' in French (a rule that allows the speaker to insert a
phoneme between words). At the grammar level, they are the number of flexions in German (for example, the same word can be written differently according to the context, like Haus, Hauses, Häuser), a high proportion of homophones (words having the same pronunciation but different spellings) amounting to 25% of the words of a French text compared to the usual 3% homophones for a German or an English text. Agglutinative languages (Finnish, Hungarian, Korean, etc.) might need a special treatment.

A new language comes with new problems to be solved. Until now we have only worked on German and English and we are just starting French, i.e. there is not much experience on multilingual recognition systems [7, 8]. The basic technology will remain the same over the languages, but language-specific implementations are unavoidable.

6. Limits of the technology

The dictation system described above works perfectly well for a single user in a domain such as radiology. The word error rate is below 10%, and if the user adapts himself to the system (this point should not be underestimated; a compromise made on both sides of conflicting positions is the most promising way to success), the word error rate will fall below 5%. But the system is speaker-dependent and domain-dependent. A speaker-independent but still domain-dependent system will have a word error rate a factor of two higher, with a factor of four in the size growth of the system, for a factor of two to three in speed reduction.

With a speaker-independent system, going further from high-quality recordings (with a microphone similar to the one used for our SP6000) to telephone quality (that could be very interesting for certain users), there is another factor of two in the word error rate. We are, then, already a factor of four higher than in our SP6000.

What about a domain-independent continuous-speech recognition system? If a speech recognizer can process complex medical reports, can it also recognize business letters, even from the same speaker, without losing a recognition performance? The answer is “no”, not without adaptation!

The DS can recognize continuous speech for large to very large vocabularies, but there is no speaker-independent general purpose continuous-speech recognition system on the market. If a DS can be implemented for radiologists or cardiologists to produce reports efficiently, the same doctor will not be able to dictate a private letter on his system until he retrain it. A continuous-speech DS actually shows best performance for narrow applications like medical reporting. The scientific community has very little experience in building
Large-vocabulary recognition

CSR systems for new and more complex applications. Moreover, a CSR system cannot switch rapidly to a new domain. We have limited our work to a few simple applications and languages (German, English), and, for these applications, the technology is at its limits.

To summarize, large-vocabulary continuous-speech recognition, worldwide, is not a general purpose technology: its introduction is vertical, in narrow technologically accessible market segments. With each new and more complex application, with each new language that will be processed, the scientific community will learn to generalize the technology. A new dimension has emerged in speech recognition: the language dimension, a no-man's land, orthogonal to the technology development dimension.

7. Summary

The last sections have shown that we have to extend our understanding of language engineering, in order to engineer language on a large scale. And this has to be done along three axes: first, along the technology axis (acoustic-phonetics, language modelling, search); second, along the application axis (legal, medical, general business etc.); third, along the language axis (English, French, German and others). Development along the last two axes cannot be done without help from the first one.

Is it possible to improve the technology? The ARPA speech recognition funding programme shows that, on two consecutive tasks, Resource Management and the Wall Street Journal, the word-error-rate has constantly decreased since 1987. And the curve over time shows we still can improve the technology! Work on longer-span language models will have to be triggered. Technological answers to language-specific problems will have to be implemented. We will have to focus on adaptation procedures, as well as on acoustic-phonetic modelling (speaker adaptation) and on language modelling (specializing on sub-domains, going to other domains).

What do we need to do this? Data, lots of data, as representative as possible, from as many different sources, applications and languages as possible. It seems to be simple, but acquiring data is very tedious work. To bypass data acquisition would be dangerous as data is the basis of our experience, data is what tells us how the world looks like.

The success of future products will be based on their capacity to be robust, i.e. to cover all problems that might occur when a naive user works with them, without technical skill or interest.

We are just beginning to tackle real-life language engineering problems. We have made the first step on a journey but we do not know how far it will lead
us. We are now facing the problem of generalizing the technology, a task for which we have very little experience.

REFERENCES


