Proceedings of the Institute of Acoustics

AUTOMATIC GENERATION OF LINGUISTIC CONSTRAINTS USING THE INSIDE-OUTSIDE ALGORITHM

K. Lari and S.J. Young

Cambridge University Engineering Department, Trumpington Street, Cambridge CB2 1PZ.

1. INTRODUCTION

Despite the recent trend towards large vocabulary speech recognition, many systems make little or no use of linguistic information when attempting to determine the identity of an utterance from acoustic data. A reason for the lack of incorporation of such knowledge may be due to the fact that their computation and representation for unrestricted English (or even for large vocabularies) is not immediately clear. The construction of simple deterministic finite state networks that work for small vocabulary systems can not be easily extended to work for large vocabularies. At the word level construction of such networks or equivalent grammar rules has proved to be very complex (Sharman, 1990) mainly because there are a vast number of "exceptions to every rule" that have to be accounted for. At the phoneme level there has been slightly more success (Church, 1987; Lee and Hon 1989), however, the networks tend to be very dense due to the large variety of pronunciations that may occur. These facts suggest that computation of rules or networks representing syntactic or phonotactic information should be computed in an *automatic* or a semi-automatic fashion.

In general two formulations are used to express linguistic constraints. These are based on the top two levels of the Chomsky Hierarchy of languages (Chomsky, 1959). The simpler of the two is to express rules in the *regular* form, however, regular grammars are generally thought to be too simple to capture details of natural languages. The second and more powerful method is to use the *context-free* formalism. The advantage of context-free grammars lies in their ability to capture embedded structures within data. Such embedded structures appear explicitly at the word level where context-free grammars are often used to model task languages (Miller and Levinson 1988; others). Recently, Church (1987) has shown that the ability to directly model embedding at the phoneme level may be effective to such an extent that it can eliminate the need for further linguistic knowledge at higher levels. Furthermore, an algorithm exists for inferring stochastic context-free grammars from a given training text. This algorithm is known as the *Inside-Outside Algorithm* (Baker 1979).

In section 2 of this paper a brief description of the Inside-Outside algorithm is presented; for a more complete discussion the interested reader is referred to Lari and Young (1990). Subsequently, a technique for incorporating linguistic knowledge into existing speech recognisers is described. Section 3 introduces a *prediction parsing* method which applies phonotactic knowledge during acoustic processing. Results are then presented for the ARM (Airborne Reconnaissance Mission) task.

2. THE INSIDE-OUTSIDE ALGORITHM

The Inside-Outside algorithm assumes that the source can be modelled by a SCFG in Chomsky normal form. The parameters of the grammar rules are therefore stored in matrices **A** and **B** with elements

$$a[i,j,k] = P(i \Rightarrow jk/G),$$
$$b[i,m] = P(i \Rightarrow m/G),$$

representing the probability of binary rules of the form $i \rightarrow jk$ (where i, j, and k are non-terminal symbols of the grammar) and terminal rules of the form $i \rightarrow m$ (where m is a terminal symbol) applied at anypoint in a derivation. Therefore, the parameters stored in the **A** matrix represent the *hidden* process while the parameters stored in the **B** matrix represent the *observable* process. Since any context-free grammar may be reduced to Chomsky normal form (Chomsky, 1959), these parameters are sufficient to describe any stochastic context-free language. Therefore, by definition the following stochastic constraint must be satisfied for all non-terminals i:

$$\sum_{j,k} a[i,j,k] + \sum_m b[i,m] = 1.$$

As with the Forward-Backward algorithm (Levinson *et al.*, 1983), the Inside-Outside algorithm addresses problems of *training* parameters from sample text and *recognition* of test sentences given the grammar. For this we define the *inner* (e) and the *outer* (f) probabilities to facilitate the analysis as follows:

$$e(s,t,i) = P(i \stackrel{*}{\Rightarrow} O(s),...,O(t)/G),$$

$$f(s,t,i) = P(S \stackrel{*}{\Rightarrow} O(1),...,O(s-1),i,O(t+1),...,O(T)/G),$$

which may be evaluated iteratively using equations:

$$e(s,t,i) = \sum_{j,k} \sum_{r=s}^{t-1} a[i,j,k]e(s,r,j)e(r+1,t,k),$$

$$f(s,t,i) = \sum_{j,k} \left[\sum_{r=1}^{s-1} f(r,t,j)a[j,k,i]e(r,s-1,k) + \sum_{r=t+1}^{T} f(s,r,j)a[j,i,k]e(t+1,r,k) \right],$$

using initial conditions:

$$e(s,s,i) = b[i,O(s)],$$

$$f(1,T,i) = \begin{cases} 1, & \text{if i=S;} \\ 0, & \text{otherwise.} \end{cases}$$

respectively. The algorithm proceeds by computing the inner probabilities in a bottom-up fashion and the outer-probabilities in a top-down fashion.

The probability of an observation sequence **O** being generated by the grammar G is

$$P(S \stackrel{*}{\Rightarrow} \mathbf{O}/G) = e(1,T,S)$$

that is, the probability of the start symbol generating the whole observation sequence from time 1 to T. This forms the recognition task.

The re-estimation formulae for the **A** and **B** parameters may be determined in a similar fashion to that of the Forward-Backward algorithm. The full derivation is shown by Lari and Young (1990) and therefore only the formulae are quoted here:

$$\hat{a}[i,j,k] = \frac{\sum_{q=1}^{Q} \frac{1}{P_q} \sum_{s=1}^{T_q-1} \sum_{t=s+1}^{T_q} \sum_{r=s}^{t-1} a[i,j,k]e_q(s,r,j)e_q(r+1,t,k)f_q(s,t,i)}{\sum_{q=1}^{Q} \frac{1}{P_q} \sum_{s=1}^{T_q} \sum_{t=s}^{T_q} e_q(s,t,i)f_q(s,t,i)},$$

$$\hat{b}[i,m] = \frac{\sum_{q=1}^{Q} \frac{1}{P_q} \sum_{t \in O(t)=m} e_q(t,t,i)f_q(t,t,i)}{\sum_{q=1}^{Q} \frac{1}{P_q} \sum_{s=1}^{T_q} \sum_{t=s}^{T_q} e_q(s,t,i)f_q(s,t,i)},$$

where Q is the total number of training sentences and P_q is the probability of generating the q^{th} sentence. The re-estimation formulae for a_{ijk} may be interpreted as the ratio of the expected number of times rule $i \rightarrow jk$ is used in a derivation, divided by the expected number of times non-terminal i is generated. Similarly b_{im} is the expected number of times non-terminal i emits terminal m, divided

by the number of times non-terminal i is generated during a derivation.

3. LANGUAGE MODELLING AT THE PHONEME LEVEL

3.1 The Prediction-Updating Process

The underlying recognition algorithm used for the work described here is the modified One-Pass algorithm which generates multiple alternatives (Young, 1984). This along with the set of trained HMM's form the pattern matcher component which automatically fits a set of models $\mathcal{M} = \mathcal{M}_1, \mathcal{M}_2, \ldots, \mathcal{M}_n$ to an unknown utterance O. Let $P(\mathcal{M}/O)$ denote the probability that the words (phonemes) corresponding to model $\mathcal{M}$ were spoken given that the utterance O was observed. The best sequence of models $\hat{\mathcal{M}}$ which describes the data then satisfies the following condition

$$P(\hat{\mathcal{M}}/O) = \max_{\mathcal{M}} P(\mathcal{M}/O).$$

The pattern matcher computes the likelihood $P(O/\mathcal{M})$, which may be used to compute $P(\mathcal{M}/O)$ through two successive applications of Bayes' Theorem (Duda and Hart, 1973):

$$P(\mathcal{M}/O) = \frac{P(O/\mathcal{M})P(\mathcal{M})}{P(O)}.$$

The above equation introduces two new terms which help to reveal the second component of the system, the *Language Model*. The quantity $P(\mathcal{M})$ represents the probability that the speaker utters the phoneme sequence represented by the underlying models $\mathcal{M} = \mathcal{M}_1, \mathcal{M}_2, \ldots, \mathcal{M}_n$. Therefore $P(\mathcal{M})$ provides a certain amount of phonotactic information. Computation of $P(\mathcal{M})$ may be performed in a variety of ways, either directly from the statistics of the transcription of speech used for training the HMM's or indirectly from stochastic grammar rules inferred from the same transcription. Both methods will be discussed in the following sections. Many researchers have assumed $P(\mathcal{M})$ and P(O) to be constant but as will be shown this simplified scenario results in a deterioration in performance. Language modelling, whether at word level or phoneme level has proved to be useful at little cost (Jelinek, 1985a; Bahl *et al.*, 1988; Lee and Hon, 1989).

In experiments performed in the following sections the above equation is expressed as follows

$$P(\mathcal{M}/O) = \frac{P(O/\mathcal{M})P(\mathcal{M})}{\sum_i P(O/\mathcal{M}_i)P(\mathcal{M}_i)}. \qquad (1)$$

the $P(\mathcal{M}_i)$ are regarded as *prior* probabilities, $P(O/\mathcal{M}_i)$ as likelihoods and $P(\mathcal{M}/O)$ as the *posterior* probability. Hence, the prior probabilities in a sense represent the process of *prediction* and posterior probabilities the process of *updating*.

The remainder of this section concerns itself with applications of stochastic grammars in determining the prior probabilities in the form of a language model. The algorithms developed for computing the prior probabilities are described in detail and their effectiveness is demonstrated for the ARM task. It will finally be shown that the grammars inferred for the task do exhibit some phonotactic properties.

3.2 The Prediction-Updating Algorithm

The approach taken in this section is to model the language generator as a SCFG in which derivation sequences form a Markov chain, the parameters of which are estimated from a large sample

of transcribed speech. Since the source is assumed to be a SCFG, the Inside-Outside algorithm is used to extract context-free rules of phonology.

Assume that a SCFG trained on some transcription of a speech database is available. Consider the situation where the pattern matcher has recognised a sequence of phonemes $p_1,\ldots,p_n$. The task of the language model may be defined as the prediction process involved in identifying the phoneme p_{n+1}. At each time frame where a prediction is required, the algorithm has to compute prior probabilities $P(\mathcal{M})$ for all models, and present the phoneme representing the model with the largest prior probability, as the predicted symbol. Once the prior probabilities are determined (the "prediction" process) equation (1) may be used to compute the posterior probabilities (the "updating" process). Computation of prior probabilities from a SCFG forms the main topic of this section. Assuming that a non-terminal j spans the phonemes $p_1,\ldots,p_n$, the idea is to proceed by locating all rules of the form $i \rightarrow jk$ and then "expanding" the right sister (in this case k) to the left until a terminal symbol is derived. All expansion probabilities can be evaluated from a set of linear equations which will be discussed shortly.

Since the prediction problem is concerned with computing the probability of occurrence of symbols at the next interval $n+1$, given a non-terminal j spanning the past n symbols, it is of fundamental interest to compute the probability l_{jk} defined as the probability of non-terminal k being immediately to the right of non-terminal j in a binary constituent rule, *i.e.* $x \rightarrow jk$. This quantity is evaluated by first recalling that

$$a_{ijk} = P(i \rightarrow jk/i, G).$$

Since all predictions are made through a binary constituent rule, define the intermediate quantity

$$\bar{a}_{ijk} = P(L = j, R = k/P = i, P \rightarrow LR, G) = \frac{a_{ijk}}{\sum_{xy} a_{ixy}}.$$

The $\bar{a}_{ijk}$'s may be interpreted as probabilities associated with 1-step derivations and therefore the daughter non-terminals j and k are viewed from above. However, during prediction, the process is regarded from the point of view of the left daughter j. Therefore define the second intermediate quantity l_{ijk} to be the probability of non-terminal k appearing in the R position of the rule $P \rightarrow LR$, given that i and j are parent and right daughter respectively,

$$l_{ijk} = P(R = k/L = j, P = i, P \rightarrow LR, G) = \frac{\bar{a}_{ijk}}{\sum_x \bar{a}_{ijx}}.$$

Now recalling that the Inside-Outside algorithm computes frequencies of usage of various rules in parsing the training data, the following probability is easily computable:

$$b_{ijk} = P(P = i, R = k/L = j, P \rightarrow LR, G) = \frac{P(i \rightarrow jk, i \text{ used})}{\sum_{xy} P(x \rightarrow jy, x \text{ used})}.$$

Hence by Bayes' theorem:

$$\begin{aligned} l_{jk} &= P(R = k/L = j, P \rightarrow LR, G) \\ &= \sum_i P(R = k/P = i, L = j, P \rightarrow LR, G).P(P = i/L = j, P \rightarrow LR, G) \\ &= \sum_i l_{ijk} \sum_x b_{ijx}. \end{aligned}$$

The probability l_{jk} may be interpreted as the probability of non-terminal k being the right sister of j. Therefore, if $\text{expL}(k,m)$ represents the probability of non-terminal k leading to terminal m through a series of "left-derivations", then

$$pred(j,m) = \sum_k l_{jk}.expL(k,m),$$

where pred(j,m) is interpreted as the probability of observing terminal symbol m given that j spans the already observed data. The expansion probabilities expL's may be computed from a set of linear equations whose coefficients are the parameters of the SCFG:

$$expL(i,m) = \sum_j \sum_k a_{ijk} expL(j,m) + b_{im},$$

where i, j, k are non-terminals and m is a terminal symbol.

So far, the prediction process has been regarded as that of computing the probability of symbol occurrence given that non-terminal j spans some part of the data already observed. The probability associated with a non-terminal spanning a portion of the data may easily be computed from the inner probabilities. Therefore

$$e(s,n,j)pred(j,m) = P(j \stackrel{*}{\Rightarrow} p_s, ..., p_n, p_{n+1} = m/G),$$

is the probability of j spanning $p_s, \ldots, p_n$ and leading to $p_{n+1} = m$. This estimate is based on the past $n - s + 1$ symbols. It is not obvious how much of the data should be used in the prediction process, intuition says that all the data must be used, but in practice the recent data contributes most to the probability. However, limiting the data too severely is unjustified since a non-terminal can in theory span a large portion of the data string. For the experiments described below, all the data is considered by summing over all non-terminals at all intervals:

$$P(\mathcal{M}_m) = \frac{\sum_s \sum_j e(s,n,j)pred(j,m)}{\sum_x \sum_s \sum_j e(s,n,j)pred(j,x)},$$

where j is a non-terminal symbol, m a terminal symbol, $s = 1, \ldots, n$ and $\mathcal{M}_m$ represents the model representing phoneme m. The prior probabilities computed in this fashion may be combined with the pattern matcher produced likelihoods to compute the posterior scores.

Since the input speech is analysed on a frame by frame basis, at each frame all possible phonemes are considered. A minimum phoneme duration of 2 frames (for short bursts and plosives) and a maximum duration of 30 frames (for vowels) is assumed. At each time frame the space between the last 30 and the last 2 frames is searched to determine the best phoneme match along with the optimal boundary. The posterior probability for each phoneme is computed by combining its likelihood with the previously computed prior probabilities, using (1). However, (1) may not be used directly because the likelihoods $P(\mathbf{O}/\mathcal{M})$ computed by the pattern matcher are very small and therefore only their logarithms are stored. Hence, (1) may only be evaluated using log arithmetic (Kingsbury and Rayner, 1971).

3.3 Incorporation of Phonotactic Knowledge

The language modelling task described in the previous section was tested on the ARM database, which was kindly supplied by Martin Russell of the Speech Research Unit (SRU) at Royal Signals and Radar Establishment (RSRE). The aim of the ARM project is accurate recognition of continuously spoken airborne reconnaissance reports using a speech recognition system based on phoneme level continuous density hidden Markov models. The details of the ARM project is described by Russell *et al.* (1990). The database (or at least the parts used in the experiments below) consists of 45 reports, 36 of which were used for training and 9 for testing purposes. Each report contains 6 to 10 sentences, which have been phonetically transcribed. The vocabulary size is about 530 words. The SCFG's inferred by the Inside-Outside algorithm were allowed to use 9 non-terminals and 45 terminals including a silence symbol. The grammar was trained on the transcription of the 36 reports used for

training the HMM's. However, the transcription could not be used directly because sentences varied in length up to 120 phonemes with an average of about 60. Since the Inside-Outside algorithm is cubic in complexity in terms of both the number of non-terminals and the input string length, a computational constraint is imposed. Therefore, each sentence of the 36 reports were segmented in various ways. Grammars were trained based on single transcribed words (in the linguistic sense), on pairs of transcribed words and triplets of transcribed words. The latter two incorporate some boundary information. The effects of pre-training is also considered. The results for the above test are tabulated in table 1.

Data	Training	*Performance*(%)
pairs	pre-trained	55.2
pairs	random	55.0
single	pre-trained	54.8
single	random	54.3
triplets	random	53.5

Table 1 - Effects of SCFG Language Modelling

Language Model	Performance (%)
SCFG	55.2
SRG	52.3
bigram	52.5
None	43.5

Table 2 - Comparison with other Language Models

Language Model	Performance (%)
SCFG (7 non-terminals)	54.8
SCFG (9 non-terminals)	55.2
SCFG (11 non-terminals)	54.6

Table 3 - Effects of Grammar Size on Performance

The performance is computed as follows:

% Correct= 100 - % Deletion - % Substitution - % Insertion.

It is interesting to note that the pre-trained SCFG trained on pairs of transcribed words improves the baseline performance by 12% whereas language models such as SRG's (or bigrams) only help 9%. The pre-trained grammars result in as good a recognition rate as those started from random initial models, taking 35% less iterations. It appears that word boundary information is useful in the sense that the SCFG's on pairs did slightly better than the SCFG's trained on single words, but it is surprising to observe that the SCFG trained on triplets does worse than those trained on pairs. This

is assumed to be due to undertraining (and perhaps too much boundary information is not helpful, in the sense that between-word phonotactics and within-word phonotactics are confused). Finally, to investigate the effects of grammar size (the number of non-terminals) the data set consisting of pairs of transcribed words was used to train different size pre-trained grammars. The results obtained are shown in table 3. This suggests that 9 non-terminal symbols is optimal for this particular task. Reduction of the number of non-terminals to 7 reduces the performance slightly but on average each iteration takes 40% less time.

Unfortunately, the automatically inferred stochastic context-free grammars do not exhibit any striking rules of phonology or phonotactics. This is partially due to the fact that rules in Chomsky normal form are too "simple" in structure to reveal the inner acts of derivations. However, their generative capacity may be examined by observing the sentences which they produce. In order to simplify matters a gross approximation is made by clustering all terminals into two broad categories, vowels and consonants. This grammar was then used to generate 100 random sentences (or strings). These sentences were compared with standard consonant-vowel combination that occur in English words, as described by Gimson (1975). 94% of these sentences formed valid consonant-vowel sequences. The remaining 6% were sentences which were either composed of only consonants or sentences with five consonants in a row. For the grammar trained on the same data, but with randomised initial parameters, 89% proved to be valid. The remaining 11% were sentences composed of only consonants or only 2 vowels. In one case four consonants appeared prevelar and three afterwards, which disagrees with Kiparsky's (1982) phonotactic constraint on length. Therefore it appears that SCFG's make a reasonable attempt in modelling phonotactics in the "broad" sense, but, it in the "narrow" sense, it is difficult to evaluate their effectiveness. This is mainly due to insufficient data being available to capture the details of the phonotactics.

	Training Data		Pre-Trained		Random Init.	
Pairs	Initial	Final	Initial	Final	Initial	Final
CV	1324	684	1297	596	1131	469
VC	343	852	342	850	339	805
CC	292	424	319	489	409	594
VV	1	0	2	25	81	92

Table 4 - Initial and Final pair distributions for the training data, sentences generated by the pre-trained SCFG and the random initial valued SCFG

Since there are many different CV combinations in English words, it is inaccurate to compare the sentence distributions of the SCFG's directly, but a test in the spirit of O'Connor and Trim (1953) is viable. This test simply examines all the sentences in the training set and notes all the initial and final pairs, using C or V, respectively, to represent any phoneme in these classes. The trained SCFG(s) are then used to generate the same number of sentences (or words) and their initial and final pairs are also noted. These frequencies are displayed in table 4. By comparison, the distribution of initial and final pairs for the trained grammars are similar to that of the original data the grammars were trained on. The pre-trained grammar is closer to the real distribution than the grammar initialised with random parameters. There happened to be no words ending with two vowels and only one word beginning with two vowels, but the trained grammars (especially the one with random initial parameters) make a generalisation and allow two consecutive vowels in those positions. As it happens there are many words ending with two vowels, it just turned out that none occurs in the ARM task. The fact that V units stand side by side less frequently than C units makes

a statement of distribution in terms of discrete units with V as a central element and C as marginal element more economical than one in which C were taken as central and V as marginal (O'Connor and Trim, 1953).

4. CONCLUSIONS

It may be concluded that a SCFG is more capable of supplying the *a priori* knowledge than an HMM, for the ARM task. Although the difference in recognition rates observed empirically is small, it is of statistically significant. It was noted that although the rules inferred do not exhibit any striking phonotactic constraints, it is beneficial to analyse the grammar through the sentence distribution. It was then shown that 94% of the sentences generated by the pre-trained SCFG formed a valid consonant vowel combination (as outlined by Kiparsky (1982)). Furthermore, a test in the spirit of O'Connor and Trim (1953) showed that the initial and final consonant vowel combinations for sentences generated by the SCFG's were very close to the actual distribution.

5. REFERENCES

Baker J.K. (1979). Trainable Grammars For Speech Recognition. Speech Communication Papers for the 97th Meeting of the Acoustical Society of America. (D.H. Klatt and J.J. Wolf, eds.) pp.547-550.

Chomsky N. (1959). On certain Formal Properties of Grammars *Information and Control*, Vol.2, pp.137-167.

Church K. (1987). Phonological Parsing in Speech Recognition, Kluwer Academic Press.

Duda R.O., Hart P.E. (1973). Pattern Classification and Scene Analysis, John Wiley and Sons.

Gimson A.C. (1980). An Introduction to the Pronunciation of English, Edward Arnold Publishers.

Jelinek F. (1985). The Development of an Experimental Discrete Dictation Recogniser, *Proceedings of the IEEE*, Vol.73, No.11, pp.1616-1624.

Jelinek F. (1985). Markov Source Modelling of Test Generation, *NATO Advanced Study Inst. Impact of Processing* Techniques on communication Martinus, Nijhoff, pp.569-598.

Kingsbury N.G., Rayner P.J.W. (1971). "Digital filtering using logarithmic arithmetic", *Electron. Lett.*, Vol.7, pp.56-58.

Kiparsky P. (1982). Explanation in Phonology, Foris-Publications Holland.

Lari K., Young S.J. (1990). The Estimation of Stochastic Context-Free Grammars using the Inside-Outside Algorithm, *Computer Speech and Language*, Vol.4 pp.35-56.

Lee K.F., Hon H.W. (1989). Speaker Independent Phone Recognition Using Hidden Markov Models, *IEEE Trans. ASSP*, Vol.27, No.11, pp.1641-1648.

Levinson S.E., Rabiner L.R., Sondhi M.M. (1983). An Introduction to the Application of the Theory of The Probabilistic Functions of a Markov Process to Automatic Speech Recognition, *BSTJ*, Vol.62, pp.1035-74.

Miller L.G., Levinson S.E. (1988). Syntactic Analysis for Large Vocabulary Speech Recognition Using a Context-Free Covering Grammar. *Proc. ICASSP*, pp.271-274.

O'Connor J.D., Trim J.L.M. (1953). Vowel, Consonant, and Syllable- A Phonological Definition, *Word*, Vol.9, No.2, pp.103-122.

Russell M.J., Ponting K.M., Peeling S.M., Browning S.R., Bridle J.S., Moore R.K., Galiano I., Howell P. (1990). The ARM Continuous Speech Recognition System, *Proc. ICASSP*, pp.

Sharman R.A. (1990). Personal Communication.

Young S.J. (1984). Generating Multiple Solutions From Connected Word DP Recognition Algorithms, *Proc. Inst. Acoust.*, Vol.6, Part 4, pp.351-354.

Proceedings of the Institute of Acoustics

SPEECH AND LANGUAGE PROCESSING AT BRITISH TELECOM RESEARCH LABORATORIES

R.D. Johnston and R.M. Brooks

British Telecom Research Laboratories, Martlesham Heath, Suffolk, U.K.

ABSTRACT

This paper provides an overview of the speech and language processing activities at British Telecom Research Laboratories. We outline the main applications which provide the 'pull' for the current research topics. Some of the immediate technical problems which we need to overcome are highlighted and the main development activities described.

1.The Speech and Language division

The Speech and Language Division at British Telecom Research Laboratories was formed in 1983. This reorganisation brought together a number of research, development and systems engineering groups who until then were working independently on speech related applications. This provided a 'critical mass' of speech researchers and meant that it was possible to extend work beyond the traditional area of speech transmission and work started on recognition and synthesis. The division has grown considerably since then and while 'speech' remains the main theme of the division we also have important projects in the areas of natural language processing and speech translation.

2. The Applications focus

2.1 Current application areas.

As British Telecom is first and foremost a telecommunications company our activities are directed toward applying the technology to the telephone network.
Typical of the services which employ speech technology are:

- voice messaging
- finance
  - . telephone banking
  - . stocks and shares
  - . insurance quotations
  - . credit card transactions
- entertainment
  - . betting
  - . horoscopes
  - . games
  - . competitions
- information services
  - . timetable
  - . yellow pages
- telemarketing
  - . promotions
  - . market research
- teleshopping/reservations
  - . theatres
  - . airlines
- field operations
  - . data entry and retrieval
  - . field personnel job dispatch
  - . voice access to electronic mail
- automatic operator
  - . network services
  - . customer premises

Although these network based services represent the bulk of our applications, research and development is also directed at enhancing terminal equipment for use at customer premises. This includes voice controlled telephones, hands-free cellphones and for improving the services available to visually handicapped employees and customers.

2.2 Core Technologies

The above applications require a mix of technologies. These are developed as individual components so that they can be engineered into different systems. The most significant of these core technologies are:

. Speech coding

. Speaker independent recognition

. High quality synthesis

. Speaker verification

. Dialogue Design

Although the algorithms and techniques developed may be used directly in applications most of them are also hosted by the PC based BT Speechcard [1].

It is the combination of the core technologies, coupled to a knowledge of the constraints of the telephone network which is the key to the development of speech interactive services.

3. Speech Coding

3.1 Current Practice

Pulse Code Modulation (PCM) is used to transmit telephone speech over the main network. Band limited speech (300Hz - 3.4kHz) is sampled 8000 times per second with a precision of 8 bits/sample. Each sample is instantaneously companded using a psuedo-logarithmic compression characteristic which provides a subjective performance equivalent similar to that of 13 bits linear coding. Worldwide two variants of companding laws co-exist: "A-law" which is used mainly in Europe and "u-law" in the USA. The differences are primarily in the position of the 'knee' of the characteristic and in the treatment of zero crossings. Where necessary, one standard can be digitally recoded to the other with only a slight deterioration in quality.

In either case the net result is that analogue speech is transformed into a 64 kbit/s digital stream and this forms the basic building block for all switching purposes. The streams may subsequently be exploited either by combining several to provide higher bit rate systems (for music quality transmission) or a single stream may be subdivided to provide multiple speech channels.

3.2 Present trends

Over recent years, not only has this 64kbit/s standard become firmly established but with the widescale deployment of low cost and wide bandwidth optical fibre systems the need for sophisticated coding schemes within the main network has diminished.

At the same time, however, this has been balanced by an increased demand for robust low bit rate coders in the peripheral network. Consequently cellular systems, specialist communication systems and speech storage systems have all been growth areas where novel coding methods been exploited.

Today the main advantages of the technology are to be found where bandwidth is still at a premium – either because there is a need to squeeze as many channels as possible into a limited radio spectrum (e.g. digital cellular telephones) or because radiated power is limited by dimensional constraints (e.g. small aerials on ships or aircraft).

3.3 Low bit rate Coding

Below 64 kbit/s the first natural subdivision is 32kbit/s. Recently, after several years of development and evaluation in which BTRL took an active part, a standardised form of Adaptive Differential Pulse Code Modulation (ADPCM) was recommended by CCITT [2]. This is now adopted as the standard for compressing two speech channels into a single 64kbit/s stream and has been applied in satellite communications and some value added services.

Our current activities therefore are mainly concerned with using the range from 16 kbit/s down to 2kbit/s. At 16kbit/s rate a sub-band codec has been implemented on a Motorola DSP 56000. This is incorporated in the BT Speechcard where it not only provides savings in storage space, but realises a significant systems engineering advantage by reducing the data transfer rate required within multiply configured networks.

At the next level of bit rate reduction, a consortium including British Telecom, British Airways and Racal Decca has recently developed the 'Skyphone' service which is marketted by British Telecom International[3]. This uses a 9.6 kbit/s codec based on multipulse linear predictive coding and implemented on a floating point AT&T DSP32.

At lower bit rates, coding techniques merge with those of speech analysis and synthesis. Using formant analysis and synthesis (copy synthesis), it is possible to provide high quality speech at 2kbit/s but at present this process is much slower than real-time. Even if the real-time problem were to be eliminated using faster processors, the approach is unsuitable for transmission purposes because the inherent 'block' processing delays are longer than can be tolerated in most two-way communications systems. These techniques are therefore primarily for use where recorded messages are required.

3.5 High quality speech at 64 kbit/s

Although normal telephony is confined to 300 – 3.4kHz, alternative coding methods may be used to increase this while remaining within the 64kbit/s limit. Recent advances, notably the use of low time delay sub-band coding, permits bandwidths of up to 7kHz and offers the potential of high fidelity telephony over the ISDN and in digital PBX environments.

Other applications include improved audio/visual conferencing, broadcast and entertainment services. Research in this area addresses not only the key problems of how to optimise the performance of coders, but also examines how standards may be formed and how such systems could interwork with established coding schemes.

3.6 Music Coding

SPEECH AND LANGUAGE PROCESSING AT BTRL

For the transmission of high quality music, a bandwidth of at least 15kHz is normally considered to be the minimum required. Not only does music require a greater bandwidth, but the human ear is less tolerant of distortion in music than in speech and more precision most be introduced during the digitisation phase.

Using sub-band coding methods, BTRL have developed a high quality music codec which encodes music sampled at 15 kHz into two 64 kbit/s channels. This represents a bit rate several times lower than that currently used in many applications.

4. Speech Synthesis

4.1 The applications focus
The immediate goal is to provide high quality speech output for applications such as directory services and recorded messages. In such cases a totally flexible Text-To-Speech (TTS) system, although desirable, is not essential and so activities are concentrated mainly on providing high quality speech for restricted vocabularies (e.g. strings of digits, amounts of money etc.). Applications already in use include Directory Automated Services (DAS), TOPAZ cellphone, Phonepoint, changed number intercept services and voice guidance for visually handicapped operators.

4.2 Text-To-Speech
Although still considered to be of lower quality than acceptable for widespread deployment in the public network, TTS synthesis is used in a number of specialised areas within BT. The 'Caesar' system which provides spoken reports of line faults for field engineers is described in Brooks[4].

Current research spans both formant and waveform based synthesis techniques and is coupled to advanced work on non-linear phonology and natural language processing. The capacity of neural nets to provide the non-linear mappings between text and phonetic level is also being explored.

4.3 Concatenation
Although less sophisticated than TTS, concatenated speech provides a flexible, yet high quality solution for speech response systems. For many envisaged applications a small number of words, frequently the digits, is sufficient to provide the voice response. However the quality must be high and users must experience little difficulty in listening to them. This is particularly important in applications such as directory inquiries and changed number intercept. Only slightly larger vocabularies are required for reading out quantities of money (e.g. billing information) or for providing simple messages.

The techniques used in concatenation range from those normally associated with copy synthesis to those of low bit rate coding and the best solutions tend to draw from both areas.

5. Speech recognition

5.1 Applications background
The first product provided by BT which exploited speech recognition was the TOPAZ cellphone[5]. This offered "hands free repertory dialling" for use in mobile applications and

was developed in 1986. It provided speaker dependent recognition of up to 100 utterances and the speech recogniser, speech feedback module, dialogue controller and interface to the cellphone were integrated into a single module using surface mount technology.

More recently the BT Speechcard (PC based) has provided the main platform for hosting the technology and the first speaker independent recogniser with a fully BABT approved interface was developed in 1989.

5.2 Telephony based recognition

Since 1986 our main development activities have been focussed on speaker independent recognition over the telephone network and are directed at addressing the applications areas described earlier. To a considerable extent this distinguishes the technology from much of that undertaken in non-telephony speech laboratories.

For example, much contemporary research into speech recognition is directed towards increasing vocabulary size and in exploring how higher level linguistic knowledge may be used to resolve ambiguities. In such cases the acoustic environment is usually both reasonably quiet and stable and typically high quality noise cancelling microphones are used to provide a clean signal.

In telephony based applications the environment is very different. In particular the signal levels may vary considerably between calls (30dB would not be unusual), the frequency shape (spectral 'tilt') of the lines may alter by +/- 12dB according to length and loading characteristics, up to 30% total harmonic distortion may be present from older carbon granule microphones and, in the early stages at least, users will not be familiar with the technology.

The difficult and highly variable signal conditions, enormous (and unpredictable) user populations and the fact that most applications can be served by the provision of small vocabulary systems means that there are few opportunities to exploit either adaptive or 'top down' techniques. Indeed in almost all cases the key requirement is robust recognition based on the acoustic signal alone.

It is not surprising therefore that our activities here have been principally concerned with optimising the performance of the digits and small application specific vocabularies spoken over the telephone line.

5.3 Recognition research

Current research activities span a number of other areas. In particular the development of a robust speaker independent connected digit and alphabet recogniser is well advanced and the performances of various subword recognition techniques are being assessed 'head-to-head' against existing whole word methods. For applications such as operator services, and ultimately directory enquiries, the need for large vocabulary recognition (possibly incorporating the recognition of spelt words) has been identified as being of particular importance.

6. Verification

6.1 Background

Humans are excellent at recognising words even in high ambient noise conditions. Indeed, given a list of many thousands of words most of us would not only be able to identify (and spell) most of them but we would also be able to point out homophones and words which were likely to be confusible if the accent of the talker was not known. Our ability to

identify a talker from a single word is much less impressive. Not only do we have difficulty in recognising a voice – but we have difficulty in associating it with names. Not only does common experience suggest this, but it is borne out from quantitative evidence, particularly that arising from forensic analysis[6].

What is not yet established is the extent to which this limit is set by the content of the signal. Perhaps we discriminate poorly between talkers because it is of little benefit to us to do so. Alternatively it may be that the limit is set by the information contained within the signal itself. To help resolve this basic issue, which clearly sets an upper limit to the performance which may be achieved, we are carrying out both theoretical and experimental investigations to determine how much speaker dependent information is contained within the signal.

6.2 Techniques being explored

On the applications front we are investigating both 'password' based techniques and 'text independent' methods: again based on the comparative measurement of performance over telephone networks.

Like recognition, much of this work is concerned with evaluating different algorithms and this process relies on having representative databases which span most of the factors which are likely to affect performance. One database has now been collected over a period of two and a half years to enable us to assess to what extent verification reliability is affected by changes in voices over time. Other problem specific databases are used to compare the sensitivity of verification algorithms against various talker/telephone/line factors.

This work is backed up with systems research exploring how the basic 'acoustic' verification can be enhanced using multiple utterances, background knowledge and dialogues.

6.3 Applications

Almost all applications which involve access to information can benefit from some form of user verification. In financially based services such as telebanking, the need for security is paramount, but there are a number of other areas where verification provides additional functionality. Voice messaging applications, interactive telephone answering machines and all the applications associated with 'Follow me' type calls and remote database access are obvious candidates.

7. Speech Technology Assessment

7.1 Established methods

The comparative assessment of speech technology forms such a significant part of our activities that it warrants special discussion.

Speech quality assessment has been carried out for over 50 years within BT (formerly the Post Office) as it was, and remains, central to the business. Methods for assessing speech quality over transmission links are well established and standard instrumentation is available for measuring the main characteristics of speech (e.g. loudness and activity) which are necessary to conduct such evaluations reliably. It has therefore been our policy to build on these to evaluate recognition, synthesis and verification technologies. This is not only because these techniques well-established, but they are tightly linked with operational parts of the business.

SPEECH AND LANGUAGE PROCESSING AT BTRL

As is well known, the original standard for speech quality was defined using the concept of the 'one metre air gap'[7]. Under this definition 'reference' speech was defined as that received at the ear when spoken from lips one metre away. This was developed further and calibrated in such a way that other types of transmission systems could be rated against it.

7.2 Comparative assessment

The great virtue of this methodology was that it avoided the problems inherent in characterising speech. There was no need to undertake any fine analysis of the speech material: instead random, simple utterances were used, and by careful experimental design it was ensured that effects attributable to the speech material could be separated out. At the same time it provided a standard which was easily replicated and common to all speech technology. As few of the main phonetic characteristics of speech e.g. graveness, creakiness, nasalisation etc. can be quantified (although they may be described) this technique neatly sidestepped almost all of the main problems inherent in measuring speech itself and allowed us to concentrate on evaluating the technology.

The heart of all our evaluation therefore is comparative assessment. Speech recognisers are compared for accuracy – rather than measured absolutely, synthesisers are compared for quality using, for example, the Listening Effort Scale[7]. In a similar way the relative effects of different environmental conditions (e.g. bandwidth variability and noise) can be assessed.

Although much of this assessment is used to assess algorithms, we also have a programme comparing the performance of a number of commercially available speaker independent recognisers, a number of verification techniques and a number of text to speech systems. The first of these are mainly accuracy tests and the latter listening tests.

8. Databases

8.1 The need for speech databases

Although database testing has been undertaken in the past for transmission quality assessment they are no longer appropriate for the evaluation of modern telephony systems where degradations such as echo need to be included: only 'live' conversation tests can do this satisfactorily. However database testing is particularly valuable for recogniser, verifier and synthesiser assessment and we have undertaken a considerable number of experiments to do this using databases derived from clean speech.

8.2 The applications focus

Like every other aspect of assessment, the key question which must be asked before collecting a database is "What do you want to know?" For example: if the need is to know if a verification algorithm works as well using "carbon" speech as from a linear microphone then a database which contains only linear speech will be of little use. Once the hypothesis to be tested is clear, then the specification of the database is relatively straightforward and the dimensions of the database can be kept within manageable proportions.Some of these databases, and the purposes for which the were generated are discussed by Walker[8].
In future work we intend to assess the effects of hands-free telephony, speech deriving from the mobile periphery and also assess further the ability of recognisers to 'reject' spurious sounds..

SPEECH AND LANGUAGE PROCESSING AT BTRL

9.Speech interactions and dialogue

9.1 The general problem

The system dialogue is central to a successful system design: yet it is still rather an art which relies upon handcrafting by an expert designer followed by a sequence of human factors trials and further modifications. The machine must prompt in a way which is easily understood and which narrows the range of potential replies. This ensures that the speech recogniser receives the best quality of acoustic signal. In practice, of course, misrecognition will occur either because the machine misheard, or more frequently in real applications, because the user uttered an invalid word. In either case it is necessary to build in checks which ensure that only valid data is passed on. The importance of a polite, pleasant voice cannot be overstated and care must be taken to avoid repetitive loops and overlong responses.

9.2 Speech Interactive Systems

BTRL have engineered a compact 32 channel unit which is suitable for a range of speech interactive services such as telephone banking. Each unit incorporates 32 Speechcards which have the following features:

- an analogue telephone interface which operates using either DTMF or voice. Call forwarding can be achieved via loop-disconnect or DTMF out-dialling.
- Speaker independent isolated-word recognition with a 50 word vocabulary active at any instant. A fast load facility is provided so that the vocabulary may be rapidly changed within a dialogue.
- Speaker-dependent isolated word recognition with a vocabulary of up to 450 words.
- Speaker verification
- High quality waveform encoded speech at 16kbit/s.

The system uses a distributed multiprocessor architecture linked to a Local Area Network (LAN). This provides both fault tolerance and a mechanism for linking to remote host computers, file servers, system management facilities and help-desks. This flexible architecture which integrates multiple distributed units with telecommunications, host computers and management systems provides a powerful service creation platform.

10. Language

10.1 The range of language work

Language research covers a number of areas. Some of these have already been discussed where they impact upon dialogues or are closely associated with speech work (e.g. text processing for Text-to-Speech). We also have activities in the 'text' based processing domain where applications such as text summarisation, text generation and language translation represent the major opportunities.

10.2 Text summarisation
Work on text summarisation addresses two types of problem. The first is concerned with abridging a single sample of text and the second is concerned with distilling a single summary from a number of different sources which discuss the same topic (e.g. reports of the same event in different newspapers). In both cases the opportunities for exploitation are focussed on e-mail systems and we are investigating both linguistically motivated techniques and those based on statistical methods.

10.3 Text generation
Text generation is needed where it is necessary to combine different sources of information into a form which is easily absorbed by a human reader. Current work includes the transformation of tabular information into textual output and studies into how different sources may be combined to provide a coherent output. Future work is likely to become more strongly linked with dialogue with the objective of generating text in a form which elicits useful responses from humans.

10.4 Text analysis and parsing
There are two programmes of work in this area. One approach is statistical (n-gram models and neural nets) and the second due to our association with the Core Language Engine currently under development in the collaborative CLARE project under the auspices of the Stanford Research Institute.

This advanced research is not directly product related at present but is expected to feed into database access applications and provide some of the components for anticipated speech products.

11. Conclusions

This paper has given a brief overview of speech technology at BTRL. We have concentrated on the main applications and consequently several major areas have been omitted. In particular the more fundamental research activities concerned with natural language processing, signal processing and pattern recognition which provide the theoretical basis for much of the work have not been addressed fully. Nor has language translation, which provides a challenge for all the technologies been treated in any depth.

More details concerning these activities are found in Wheddon and Linggard [9].

We hope that we have shown the extent to which British Telecom is committed to evaluating, developing and exploiting the technology and indicated the wide variety of applications which we expect to see emerging over the next few years.

11. References

[1] Hunter P.J., Watts M.O: "A Speech Card for Provision of Interactive Speech Systems";*Digital Signal Processing: components and applications seminar, ERA 880386,Nov 1988.*
[2] CCITT Recommendation G721: *Blue Book Vol III, Fascicle III.4 (1988)*

[3] Boyd I. and Southcott C.B: "A Speech Codec for the Skyphone Service", *British Telecom Technical Journal, Vol 6, No 2, April 1988, pp 50-59.*
[4] Brooks R.M: " Voice Operated Databases", *SHARE European Association, Vol 1, Spring Meeting, 18-22 Apr 1988, Davos, Switzerland, pp89-105.*
[5] Forse N.J.A: "Speech Recognition for Telephony Applications", *Proceeding of the Institute of Acoustics, vol 9 (1987)*
[6] Nolan, F: The phonetic bases of speaker recognition; *Cambridge University Press (1983).*
[7] Richards, D.L: Telecommunication by Speech: *Butterworths, London, (1973).*
[8] Walker G.P: "Speech Database Collection Architecture", *this conference.*
[9] Wheddon C. and Linggard R: "Speech and Language Processing", *Chapman and Hall,(1990).*