



Towards increasing speech recognition error rates

Hervé Bourlard^{a,b,*}, Hynek Hermansky^{b,d}, Nelson Morgan^{a,c}

^a *International Computer Science Institute, Berkeley, CA, USA*

^b *Faculté Polytechnique de Mons, Mons, Belgium*

^c *University of California at Berkeley, Berkeley, CA, USA*

^d *Oregon Graduate Institute, Portland, OR, USA*

Received 21 November 1995; revised 18 January 1996

Abstract

In the field of Automatic Speech Recognition (ASR) research, it is conventional to pursue those approaches that reduce the word error rate. However, it is the authors' belief that this seemingly sensible strategy often leads to the suppression of innovation. The leading approaches to ASR have been tuned for years, effectively optimizing on test data for a local minimum in the space of available techniques. In this case, almost any sufficiently new approach will necessarily hurt the accuracy of existing systems and thus increase the error rate. However, if progress is to be made against the remaining difficult problems, new approaches will most likely be necessary. In this paper, we discuss some research directions for ASR that may not always yield an immediate and guaranteed decrease in error rate but which hold some promise for ultimately improving performance in the end applications. Issues that will be addressed in this paper include: discrimination between rival utterance models, the role of prior information in speech recognition, merging the language and acoustic models, feature extraction and temporal information, and decoding procedures reflecting human perceptual properties.

Zusammenfassung

Auf dem Forschungsgebiet der Automatischen Spracherkennung (AS) werden für gewöhnlich jene Ansätze verfolgt, die die Wortfehlerrate reduzieren. Dennoch ist es die Überzeugung der Autoren, daß diese vernünftig erscheinende Strategie oftmals die Unterdrückung von Innovationen mit sich führt. Die führenden Verfahren der AS sind über Jahre hinweg verfeinert worden und haben so auf den gegebenen Testdaten ein lokales Minimum im Rahmen der verfügbaren Techniken erreicht. In dieser Situation wird nahezu jeder substantiell neue Ansatz notwendigerweise die Genauigkeit bestehender Systeme beeinträchtigen und somit die Fehlerrate erhöhen. Jedoch sind neue Ansätze höchstwahrscheinlich notwendig, wenn bei den verbleibenden schwierigen Problemen Fortschritte erzielt werden sollen. In dieser Arbeit diskutieren wir einige Forschungsrichtungen der AS die nicht immer eine unmittelbare und garantierte Verringerung der Fehlerrate erzielen mögen, die aber letztendlich Verbesserungen der Endanwendungen möglich erscheinen lassen. In dieser Arbeit werden u.a. die folgenden Themen behandelt: Abgrenzung konkurrierender Modelle einer Äußerungen, die Rolle von Vorabinformation in der Spracherkennung, die Verschmelzung von Sprach und akustischen Modellen, Merkmalsextraktion und temporäre Information und Dekodierungsverfahren, die menschliche Wahrnehmungseigenschaften reflektieren.

* Corresponding author.

Résumé

Dans le domaine de la recherche en Reconnaissance Automatique de la Parole (RAP), il est devenu habituel de poursuivre en priorité les approches réduisant le taux d'erreurs au niveau du mot. Les auteurs émettent cependant quelques réserves à l'égard d'une telle stratégie qui, bien qu'apparemment raisonnable, conduit souvent à un manque d'innovation. En effet, d'énormes efforts ont été consentis pendant plusieurs années sur les approches aujourd'hui prédominantes en RAP et celles-ci ont été développées et testées sur des données standards de référence, convergeant donc ainsi vers un minimum local dans l'espace des techniques disponibles. Dans ce cas, il est clair que pratiquement n'importe quelle nouvelle approche suffisamment différente ne pourra pas se comparer favorablement aux systèmes existants et résultera souvent initialement en une augmentation du taux d'erreurs. D'un autre côté, il est également probable que les problèmes restant à résoudre nécessiteront de nouvelles approches. Dans ce papier, nous discutons certaines directions de recherche qui ne conduiront peut-être pas toujours à une diminution immédiate et garantie du taux d'erreurs, alors qu'elles pourraient ultimement s'avérer bénéfiques aux performances de nos systèmes. Les thèmes qui seront abordés dans ce papier concernent notamment: la discrimination entre modèles, le rôle de l'information a priori en reconnaissance de la parole, le couplage du modèle acoustique et du modèle de langage, l'extraction des caractéristiques et information temporelle, et quelques procédures de reconnaissance reflétant mieux les propriétés perceptuelles chez l'homme.

1. Introduction

“Whenever a new discovery is reported to the scientific world, they say first, ‘It is probably not true’. Thereafter, when the truth of the new proposition has been demonstrated beyond question, they say, ‘Yes, it may be true, but it is not important.’ Finally, when sufficient time has elapsed fully to evidence its importance, they say, ‘Yes, surely it is important, but it is no longer new.’...”

– Michel Eyquem Montaigne, 1533–1592 –

As noted in the abstract, the main goal of this paper is to promote the notion of “deviant” research paradigms, which, when initially explored, may increase the error rate on standard tests. This is in contrast to the standard approach in our field, in which state-of-the-art system A is gently perturbed to create system B, resulting in a relative decrease in error rate of from 1 to 10%.

This latter approach is encouraged by two related factors:

1. The need to show return on investment for industrial laboratories that have spent a significant amount of manpower (sometimes hundreds of person-years of research and development) to turn state-of-the-art technology (see next section) into working (or, preferably, selling) products.
2. The research agencies’ perspectives, which (quite justifiably) are driven more and more by potential exploitation of the funded research projects.

Of course, it is always necessary to reduce the

search space of solutions by eliminating research directions that do not appear to be promising. However, if this is done too strictly, progress can be restricted to incremental gains, and performance can converge to a local minimum from which it will be harder and harder to escape.

Permitting an initial increase in error rate can be useful, as long as three conditions are fulfilled: (1) solid theoretical or empirical motivations, (2) sound methodology (so that something can be learned from the results), and (3) deep understanding of state-of-the-art systems and of the specificity of the new approach. The title to this paper was chosen to remind us that the development of new technologies will often initially result in a significant increase in speech recognition error rates before becoming competitive with or better than the best current systems; of course many new ideas will never reach this stage, but this is true (by definition) for high risk research, though this does not obviate the need for such work. A few high-risk research directions that could have a large payoff are discussed in this paper, along with a brief description of some specific examples that the authors have encountered.

As noted previously, this paper certainly does not aim at describing and discussing all possible alternative paradigms – this may be an impossible task. In fact, it is not first and foremost a “technical” paper per se, but rather it is a paper that uses some broad notions and a few specific examples to encourage more creativity in today’s ASR research. Although

unavoidably biased towards examples from our own work, we mainly use these to illustrate the approach; certainly we know of much other creative research, and so we have tried to include many references to other interesting paradigms. Necessarily, though, these references will be incomplete due to our own ignorance. In particular, since all of the authors mostly work in the area of acoustic modeling, we will not discuss in any detail either mainstream or alternative approaches to language modeling, with the exception of some of the issues of the combination of acoustic and language modeling information.

Finally, in view of current trends in research funding (observed for example in the ESPRIT programs in Europe, and ARPA in the US), it can be difficult (though not impossible) to get support for approaches that do not give strong short-term return. This trend is understandable, and it is reasonable that a field that has been researched for so many years should be producing demonstrable and exploitable results. Nonetheless, the significance of research results has often not become obvious until many years later.

In the field of ASR research, government funding (e.g., from ARPA and ESPRIT) has had a strong impact, resulting in significant developments, pushing the technology to the point where it is commercially viable for many applications. However, this having been achieved, it is now time to go on, and to start working on basic problems that have not yet been dealt with in the current mainstream technologies. For this, the field clearly needs to be opened up again to a range of approaches, with short-term reduction of error rate being de-emphasized in favor of increasing our understanding; this will likely be required to make further breakthroughs on the scale of LPC, DTW or HMMs.

Further, more research work is needed on realistic applications so that practical problems are dealt with. For instance, there is a range of talker phenomena that is common for natural speech, such as false starts and deletions. In addition, in the United States many accents are common and yet typically research tasks have only used speech from the common American dialects. In the past much research effort has been expended on artificial databases edited to remove many of these common sources of talker variability. More recent work on natural speech query

systems (such as ATIS) and on conversational speech recognition (such as the Switchboard corpus) are steps in the right direction. However, we also need to look at human speech input for real field applications, which are beginning to become more widespread.

One caveat to the reader is that this is not an extended exposition on the technical issues discussed, but rather a tract that will (hopefully) encourage more adventurous explorations, while providing pointers to more detailed information that is already in the literature. Finally, although we tend to use HMM/ANN (see Section 3.1) formalisms in this paper (since this approach was what we used to develop and understand many of the issues addressed here), we also hope that the reader will not view this merely as another paper on this topic.

2. State-of-the-art ASR systems

In the following, the “standard ASR technology” will refer to the Hidden Markov Model (HMM) technology as used in a “state-of-the-art” ASR system. Depending on the size of the lexicon, this system will often be based on strictly left-to-right 3-state HMM phoneme models that will be concatenated to get word and sentence models. In the very best systems, diphones, triphones, and sometimes quadriphone or quintaphone models are combined (smoothed) with phoneme models to get good estimates of context-dependent phoneme models. In the case of continuous speech recognition, bi-grams, tri-grams and, more recently, quadri-grams are used to model the grammar. All of these models are trained on very large acoustic and text databases.

Since it has been observed for such systems that HMM transition probabilities per se have virtually no effect on recognition performance, explicit duration modeling is also commonly used. It seems that the best solution is often the simplest one, consisting of imposing a (trained) minimum duration (simply by duplicating states that cannot be skipped).

Discrimination has also been important, and is known to be a problem. Several solutions have been proposed and have been shown to increase performance. It is however not always clear to what extent these solutions are actually used in large vocabulary

continuous recognition systems (outside of evaluation systems).

Scanning through the proceedings of the most popular international speech recognition conferences from the last five years, we note that the topics vary from mundane developmental issues to deep and principled investigations. The former type can of course be quite important. However, we do think there has been too great an emphasis on marginal improvements, and that in fact such details can sometimes masquerade as a research issue. Some of the themes that have received considerable emphasis over recent years include:

1. Increasing the size of the training database – Of course, increasing the size of the database nearly always results in a decrease of error rate. The reason for this is that the more training data one has, the more parameters one can afford to train and, consequently, the more detailed the models can be. Paraphrasing Jordan Cohen [22], this is related to the distinction between speech description (which we usually do by increasing the number of models and parameters) and speech modeling (which we all would like to do by extracting the underlying properties and invariants of the speech signal). Therefore, in general researchers will use as much training data as they can afford (in terms of space, processing time, or other logistic difficulties). Current speech databases are sometimes large enough to train around 10^7 acoustic parameters and tri-gram grammars for a lexicon of 60K words. However, the main research issue of interest could be determining what is the best data to use, the best models to use with that data, and how to re-use the knowledge extracted from the data on a new problems, rather than just increasing the training data and noting an error rate decrease.
2. Better smoothing criteria – This topic is closely related to the previous one. With enough training data, very detailed models can be trained with very little smoothing. However, larger models with even more detail can always be considered, and so smoothing (also referred to as “regularization”) remains a deep problem. For the most part, however, most experiments in this category are minor modifications of existing approaches to smoothing, with correspondingly minor results.
3. Increasing the lexicon size – There are interesting technical problems associated with recognizing a larger vocabulary with moderate computing resources. Aside from the computing resources, there are deeper issues such as the potential explosion of acoustically confusable utterances. However, for the most part we as a group have focused on the engineering issues and not the scientific ones.
4. Speeding up the search – This is actually an example of the kind of engineering work that is encouraged by the increase in vocabulary size referred to previously. This is important work, as it allows us to (1) get real-time recognizers (even for very large lexica), (2) be able to get a performance number (i.e., global error rate) on very large (reference) test corpora, and (3) make errors much faster.
5. Robust speech recognition – While this is often proposed as a separate research topic, research and development in this area is basic to the operation of virtually any recognition system under realistic conditions, and hence is critical. Current recognizers fail badly under variable conditions that have little effect on intelligibility for human listeners. Robustness to such factors is a most interesting and important research area, but is also the most challenging one. It should address all issues related to the (very poor) robustness of current ASR systems against additive noise, convolutional noise (e.g., transfer functions of transmission channels), rapid speech, speech with spontaneous phenomena such as false starts, and other phenomena that we do not understand very well.

This challenging set of issues is sometimes partially addressed in speech analysis, or the corrections are attempted by adapting the parameters of the statistical models. An active area of research in this direction is talker adaptation. This can be performed by quickly adapting the model’s parameters (e.g., using a gradient-like procedure or by training a simple function) [1,66], or by modifying the acoustic features in an appropriate way. An interesting recently proposed technique is sometimes referred to as “vocal tract normalization” [63], which permits an iterative modification of the properties of speech analysis to opti-

mally train and use the recognizer on a given set of training and test data. Although this technique was initially proposed to handle predictable variability of the data due to the variable length of vocal tracts of different talkers, we believe that the technique is far more general and in principle may allow for handling any systematic bias (such as noise level, rate-of-speech, ...) that varies from utterance to utterance, as long as a correlate of this bias can be found in some control parameter of the analysis procedure.

However, it may be the case (though we cannot know for sure) that the changes to the baseline technology will have to be deeper to really address these problems, i.e., involving completely different kinds of signal processing techniques or even different recognition strategies. This will be further addressed later in the paper.

Automatic adaptation of the language model (mainly by adapting the bi-gram and tri-gram probabilities) to the tasks, or sub-domains, as well as to the speaker's peculiarities is also receiving a great deal of attention [27,60].

In the following sections we will discuss some research directions that are a deviation from the mainstream "state-of-the-art" technology. Support of these approaches may well be necessary, however, if we aspire for developing ASR technologies that could be applied to realistic problems.

3. Statistical estimation

Over the last twenty years (and particularly the last 10) a set of techniques based on the HMM abstraction have come to dominate speech recognition [10,58], particularly large vocabulary continuous speech recognition. One can still imagine systems that are not statistical in nature, and given the appropriate theory, such approaches should be tried.

Nonetheless, statistical approaches have proven their power in this period, and for this reason we will constrain our consideration of "alternative" approaches in this paper to those methods that are still built around a formal statistical structure. Within stochastic methods, however, there is a broad range of possibilities, while speech recognition systems have tended to develop in a "depth first fashion",

that is, optimizing a particular kind of recognition system that worked reasonably well on some problems. With some notable exceptions (e.g., [44,57,89,105]), most of these systems have been trained using a Maximum Likelihood criterion and with either discrete or mixture Gaussian estimators, and evaluated during recognition with a strictly frame-based system (that is, one in which local similarity or distance measures are evaluated once per frame, and treated as independent observations in order to combine them into a global measure to determine the best match).

However, even within the statistical framework, there have been some notable exceptions. For instance, Brown [20] and others have experimented with systems trained using a Maximum Mutual Information (MMI) criterion. Ostendorf [78], Ghitza and Sondhi [38], Goldenthal [44], and others have experimented with training segment-based models. Generalized Probabilistic Descent (GPD) [64] methods have recently been developed which also use an MMI-like criterion [known as Minimum Classification Error (MCE)] for training.

Each of these has some potential advantages over the now-standard likelihood-based HMM approach. Here we will briefly describe some of the key points in the development of hybrid HMM/neural network systems, in which a different structure is used with a different discriminant criterion. Although this approach has begun to achieve some maturity, our own experience with it began with some experiments that "successfully" increased our error rates (initially) to 130%.

3.1. Hybrid HMM / ANN systems

It has been shown that Artificial Neural Networks (ANNs) could estimate the posterior probabilities of output classes conditioned on the input [14,39,88]. As a result of this property, a number of researchers over the last seven years have built hybrid HMM/ANN systems in which ANNs have been discriminatively trained to estimate emission probabilities for HMMs [15,23,33,34,70,73,90,95]. A detailed description of hybrid HMM/ANN systems is given in [74].

Let $X = \{x_1, x_2, \dots, x_n, \dots, x_N\}$ denote the sequence of acoustic vectors associated with a speech

utterance. In a few words, the hybrid HMM/ANN approach can be summarized as follows. As in standard HMMs, each of our models is assumed to be built up from a set of C possible classes $\Omega = \{\omega_1, \dots, \omega_c, \dots, \omega_C\}$ (with which will be associated probability density functions) which are put together according to a predefined or trained topology, yielding a different HMM M_i for each word or sentence to be recognized. Each HMM is thus defined as an oriented graph M with let say K states q^k ($k = 1, \dots, K$), each of them being associated with a class $\omega(q^k) \in \Omega$.

In the case of standard HMMs, the required probability density functions $p(x_n | q^k) = p[x_n | \omega(q^k)]$ are typically assumed to be Gaussian or multi-Gaussian distributions¹. In the case of hybrid HMM/ANN systems, neural networks (with C outputs) are trained in classification mode at the acoustic frame level to generate local posterior probabilities $P[\omega_c | x_n, \Theta]$, $\forall c \in [1, C]$, where Θ denotes the parameter set on which training of the models is performed (in our case, the ANN parameters). These (local) posterior probabilities are then used, after division by the class prior probabilities $P[\omega_c]$ observed on the training set, as scaled likelihoods $p[x_n | \omega(q^k), \Theta] / p(x_n)$ in standard HMMs to compute $p(X | M, \Theta)$ [or, more precisely, $p(X | M, \Theta) / p(X | M)$, given the scaled likelihoods and assuming no correlation between successive acoustic vectors].

This kind of approach has been successfully used with Multi-Layer Perceptions (MLPs) [15] as well as Recurrent Neural Networks (RNNs) [90]. RNNs seem to present a slight advantage compared with MLPs [89], at least with regard to the number of parameters required for similar results. In spite of its simplicity, this approach offers several potential advantages over standard HMMs, including:

- Model accuracy: ANN estimation of probabilities does not require detailed assumptions about the form of the statistical distribution to be modeled, resulting in more accurate acoustic models.
- Discrimination: ANNs can easily accommodate discriminant training. Of course, as currently done

in standard HMM/ANN hybrids discrimination is only local (at the frame level). However, recent theoretical work that allows global discriminant training of hybrid systems will be briefly presented in this paper (Section 3.3).

- Context sensitivity: In the case of RNNs or if several acoustic vectors $X_{n-c}^{n+d} = \{x_{n-c}, \dots, x_n, \dots, x_{n+d}\}$ are used at the input of an MLP, local correlation of acoustic vectors can be taken into account in the probability distribution. In the case of an MLP, outputs will be estimates of $P[\omega(q^k) | X_{n-c}^{n+d}]$. For various reasons this was not feasible with standard HMMs. Two of the closest solutions implemented in HMMs were: (1) to use first and second time derivatives as additional acoustic features and (2) to consider a few adjacent frames (typically 3–5 frames in total) on which Linear Discriminant Analysis (LDA) [55] is performed to reduce the dimension of the acoustic features (see, e.g., [45]). See Section 4.2 (and, more specifically 4.2.1) for further discussion of this.
- Parsimonious use of parameters since all probability distributions are represented by the same set of shared parameters. It is also known that it is more “economical” to model boundaries between acoustic classes (i.e., posteriors) than surfaces of density functions (i.e., likelihoods).
- Flexibility: Using a neural network as the acoustic probability estimator permits the easy combination of diverse features, such as a mixture of continuous and categorical (discrete) measures.
- Complementarity: it is sometimes the case that neural networks can supply complementary information to that provided by an existing likelihood-based system. For instance, in one approach, the combination of HMMs with a neural network (referred to as “segmental neural network”) provided some improvements over the original system [8]. In that case, an N -best paradigm [93] is used to generate the N -best utterance hypotheses that are then rescored by a neural network taking complete phonetic segments into account.

More recently, it was also observed that the availability of posterior probabilities (before division by priors) allowed a more efficient pruning for large vocabulary speech recognition systems [87]. Finally,

¹ In this paper, actual probabilities will be denoted $P(\cdot)$ while probability density functions will be denoted $p(\cdot)$.

it appears that relatively simple systems can be effectively used. In particular, it has been shown that fairly simple layered “neural” structures, which we lately have termed Big Dumb Neural Networks (BDNNs), can be used for this purpose. This name was chosen to match our discovery that large, relatively simple structures were sufficient for the statistical estimation purpose in speech recognition. These systems employ a single very large hidden layer (with 200 to 4000 hidden units, depending on the task), and are simply trained with online back-propagation with a relative entropy error criterion [15].

However, to be fair, we should note that hybrid HMM/ANN approaches typically require more computational resources for training than do the traditional systems. This is probably one of the main reasons (in addition to the “non-mainstream” character of the research) why there were initially very few successful applications of neural networks to large-scale ASR problems.

Many (relatively simple) speech recognition systems based on this hybrid HMM/ANN approach, have been proved, on controlled tests, to be both effective in terms of accuracy (comparable or better than equivalent state-of-the-art systems) and efficient in terms of CPU and memory run-time requirements (see, e.g., [70,89]). More recently, such a system (ABBOT from Cambridge University, see, e.g., [54]) has been evaluated under both the North American ARPA program and the European LRE SQALE project (20,000 word vocabulary, speaker independent continuous speech recognition). In the preliminary results of the SQALE evaluation (reported in [96]) the system was found to perform slightly better than any other leading European system and required an order of magnitude less CPU resources to complete the test. Another striking result is that the acoustic models for this system used several hundred thousand parameters (around 500,000 for ABBOT) while the corresponding models for the competing systems used millions of parameters (around 10^7 as mentioned in the introduction). While the language models for all the systems used a comparable number of parameters (also millions), this still had a significant effect on the practical implementation, and also was an experimental confirmation of the succinctness of neural network probability estimators.

This result is a far cry from the kind of results we

and others achieved in our early investigations of this class of technique. For instance, on our first attempts, we only succeeded to boost the recognition error rate (on a standard assessment database) from about 10% to 130%. It took us a number of years to get the new system to a good level of performance. Still, we feel that the investment of our time was a good one, since we now have a set of tools (in addition to the standard ones) that has many interesting features (as discussed above) and allows for many possible extensions; some of these are currently being investigated and are briefly described in some of the next sections.

We close this discussion by noting that in general it can take many years for a paradigm to be widely accepted. Probably the most obvious example of this in speech recognition is the use of HMMs for speech recognition, as pioneered by Jelinek [58] and Baker [10]. This class of approaches did not really achieve wide acceptance until the mid 1980s. It is important that researchers to some extent listen to their own internal sense of confidence, and not determine a research path by taking a poll of Wise Researchers.

3.2. Discriminant approaches in HMMs

As noted previously, there has been much recent interest in discriminant criteria for HMM training, such as MMI [20] or GPD/MCE [64]. Of course, linear discriminant approaches have been in use for many years, and have been applied successfully to problems such as connected digit recognition [30,55]. Corrective training (first proposed by Bahl et al. [9]) has been applied in the past to specifically deal with the misclassification problem (due to the lack of discrimination of the usual HMM training criterion).

In our own work, hybrid HMM/ANN systems as we currently implement them are a modification of an early theory we initially called discriminant HMMs, which was initially developed to estimate global posterior probabilities $P(M | X)$. In [15], it is indeed shown that the global posterior probability $P(M | X, \Theta)$ of a Markov model M given an acoustic vector sequence X (and parameter set θ) could be expressed as

$$P(M | X, \Theta) = \sum_{\forall Q_j} P(M, q_{1j}, q_{2j}, \dots, q_{Nj} | X, \Theta), \quad (1)$$

in which “ $\forall Q_j$ ” represents all possible (legal) state sequences in M , q_{nj} the specific state visited at time n for path Q_j . Each term in (1) can further be decomposed into

$$P(q_{1j}, q_{2j}, \dots, q_{Nj} | X, \Theta) \\ \times P(M | q_{1j}, q_{2j}, \dots, q_{Nj}, X, \Theta) \quad (2)$$

and, under the assumptions stated in [15], we have

$$P(q_{1j}, q_{2j}, \dots, q_{Nj} | X, \Theta) \\ = \prod_{i=1}^N P(q_{nj} | q_{n-1,j}, x_n, \Theta). \quad (3)$$

The second factor in (2) can be considered independent of the acoustic sequence X (since the state sequence is given) and then represents the contribution of the language model. As discussed in [16], depending on what we encode into the acoustic models, this latter factor will represent phonological, lexical and/or syntactical information.

Discriminant HMMs are thus described in terms of Conditional Transition Probabilities $P(q_n^l | q_{n-1}^k, x_n)$ for all possible state pairs allowed by M . Notation q_n^l stands for the specific state q^l of M hypothesized at time n and $P(q_n^l | q_{n-1}^k, x_n)$ denotes the probability of transition from the k th to the l th state at time n given the observation of acoustic data chunk x_n . As with traditional hybrid HMM/ANN systems, these conditional transition probabilities can be estimated by an ANN with K output units and in which the acoustic input x_n is complemented by a set of additional input units representing the state q^l hypothesized at the previous time step $n - 1$. Of course, the network will then have to be estimated for all $[\omega(q_{n-1}^k), \omega(q_n^l)]$ -pairs allowed by the HMMs. The conditional transition probabilities are thus functions of Θ and are given by $P[\omega(q^l) | \omega(q^k), x_n, \Theta]$.

Of course, as done with previous hybrid HMM/ANN systems, x_n will usually be replaced by $X_{n-c}^{n+d} = \{x_{n-c}, \dots, x_n, \dots, x_{n+d}\}$ to take some acoustic context into account.

The reader should note that, in contrast to traditional HMMs, Discriminant HMMs have inputs (the acoustic vectors) rather than outputs. Thus, the states are acceptors rather than emitters, and represent a recognition or detection mechanism.

Facing numerous problems, this approach was initially simplified by (1) disregarding the previous state in the conditional and (2) dividing the local posteriors by the class priors to get scaled likelihoods for use in standard HMMs, which led to the hybrid HMM/ANN system described in Section 3.1. The generally accepted reason for this division by class priors is briefly discussed in Section 3.5.1.

As noted above, we modified the original theory in order to develop practical HMM/ANN systems for some years. Recently, however, we have learned how to train Discriminant HMM systems in a way that is much closer to this original theory (and in fact does not require the division by class priors). The new approach, referred to as Recursive Estimation and Maximization of A Posterior probabilities (REMAP), is briefly described in the next section.

3.3. REMAP for Discriminant HMMs

3.3.1. Motivations

The motivations of REMAP are twofold. Firstly, we want to define a training algorithm for Discriminant HMM that directly optimizes the parameter set Θ according to the Maximum A Posteriori (MAP) criterion, i.e., maximizing $P(M | X, \Theta)$ if M is the correct HMM associated with X and, consequently, minimizing the actual error rate (cross-validation will be used to guarantee that minimization of the error rate does not happen only on the training data but also on an independent test set). This algorithm computes successive estimates of new (local) posterior probabilities that are used as targets for ANN training, followed by a standard ANN training, to guarantee an iterative increase of the global posteriors. Estimation of the new ANN targets involves “forward” and “backward” recurrences that are reminiscent of the Forward-Backward (Baum-Welch) algorithm [58,69,86], a particular instance of the Expectation Maximization (EM) algorithm [28] as applied to HMMs.

The second motivation of REMAP is to compute smooth transition probabilities. Indeed, Discriminant HMMs as well as other methods (e.g., SPAM – see Section 5.2) use conditional transition probabilities as the key building block for acoustic recognition. It is, however, well known that estimating transitions accurately is a difficult problem. In our previous hybrid systems, the targets used for ANN training

are typically given by the best segmentation resulting from a Viterbi alignment. This procedure thus yields rigid transition targets, which may not be optimal in the case of training (and testing!) of conditional transition probabilities.

3.3.2. Problem formulation

Global MAP training of Discriminant HMMs should find the optimal parameter set Θ maximizing

$$\prod_{i=1}^I P(M_i | X_i, \Theta), \quad (4)$$

in which M_i represents the Markov model associated with each training utterance X_i , with $i = 1, \dots, I$.

Although, in principle, we could use a generalized back-propagation-like gradient procedure in Θ to maximize (4) (see, e.g., [11]), an EM-like algorithm should have better convergence properties, and would preserve the statistical interpretation of the ANN outputs. In this case, “full” MAP training of transition-based HMM/ANN hybrids² requires a solution to the following problem: given a trained ANN at iteration t providing a parameter set Θ^t and, consequently, estimates of $P[\omega(q_n^l) | x_n, \omega(q_{n-1}^k), \Theta^t]$, how can we determine new ANN targets that:

1. Will be smooth estimates of conditional transition probabilities, \forall possible (k, l) state transition pairs in M and for each possible time frame $n \in [1, N]$ (where N is the utterance length).
2. When training the ANN for iteration $t + 1$, will lead to new estimates of Θ^{t+1} and new

$$P[\omega(q_n^l) | x_n, \omega(q_{n-1}^k), \Theta^{t+1}]$$

that are guaranteed to incrementally increase (4)?

In [16], we prove that a re-estimate of ANN targets that guarantee convergence to a local maximum of (4) is given by³

$$\begin{aligned} P^* [\omega(q_n^l) | x_n, \omega(q_{n-1}^k)] \\ = P(q_n^l | X, q_{n-1}^k, \Theta^t, M), \end{aligned} \quad (5)$$

² Although REMAP is also valid for “standard” HMM/ANN hybrids, we decided to focus on transition-based probabilities as required by Discriminant HMMs and a new model called SPAM.

³ In the following, we consider only one particular training sequence X associated with one particular model M . It is, however, easy to see that all of our conclusions remain valid for the case of several training sequences X_i , $i = 1, \dots, I$.

which means that the new ANN target associated with x_n and a specific transition $q^k \rightarrow q^l$ has to be calculated as the probability of that specific transition conditioned on the whole training sentence X and the associated model M .

In [16], we further prove that alternating ANN target estimation (the “estimation” step) and ANN training (the “maximization” step) is guaranteed to incrementally increase (4) over t ⁴; we also provide efficient forward and backward-like recurrences to compute (5).

3.4. Discussion and results

Of course, a wide range of discriminant approaches (e.g., MMI [20], GPD/MCE [64] – see [16] for a comparative discussion of these) to speech recognition have been studied by researchers. A significant difficulty that has remained in applying these approaches to continuous speech recognition has been the requirement to run computationally intensive algorithms on all of the rival sentences. Since this is not generally feasible, compromises must always be made in practice. For instance, estimates for all rival sentences can be derived from a list of the “ N -best” utterance hypotheses, or by using an ergodic model containing all possible phonemes. This is not required with REMAP.

Although much work is still required to optimize the practical heuristics for this method, preliminary results [17] show some recognition improvement on isolated digits and continuous natural numbers.

We note here that REMAP is somewhat related to the Input–Output HMM (IOHMM) [12] that can also yield discriminant global posterior based training.

In the next sections we discuss a few of the many open issues in statistical estimation for speech recognition. Each of these areas needs work for the increased understanding that will be necessary for

⁴ Note here that one “iteration” does not stand for one iteration of the ANN training but for one estimation-maximization iteration for which a complete ANN training will be required.

significantly better performance for very hard problems in this area.

3.5. Open issues

In the following, we discuss a few issues that we believe are still poorly understood in ASR systems, both for “standard” HMMs or hybrid HMM/ANN systems (although HMM/ANN systems helped us to identify the potential problems).

3.5.1. Likelihoods and priors

In standard (likelihood-based) HMM acoustic training, the prior probabilities of models are not used during likelihood training. [The reader should however note here that in this paper, and especially in this section, we are using the word “priors” to refer to the prior distributions for models and their topologies, rather than the common usage of the word as a probability distribution over model parameter values.] In other words, the language model parameters are trained independently of the acoustic model parameters or are fixed by a priori knowledge. This results from the application of Bayes’ rule to the statistical formulation of the speech recognition problem (as originally stated in [58,59]). In this approach, the optimal criterion $P(M | X, \Theta)$ is expressed as

$$P(M | X, \Theta) = \frac{p(X | M, \Theta) P(M | \Theta)}{p(X | \Theta)}, \quad (6)$$

which separates the probability estimation process into two parts: (1) the Language Modeling $P(M | \Theta)$ and (2) the Acoustic Modeling $p(X | M, \Theta)$, referred to as Acoustic Likelihood.

The dependence of the denominator of (6) on the parameter set Θ is irrelevant during recognition, when the parameters are (typically) fixed. During training, this factor is also typically ignored. While even during training there is no explicit dependence on M for this denominator, in fact it will change as the parameters are modified, so that the modification of the parameters for a training sample may improve the likelihood of the correct model and yet also improve the likelihood of incorrect models; note that the denominator may also be viewed as the sum over

all models of the product of each acoustic likelihoods and its corresponding model priors. Thus using the likelihood criterion for training acoustic parameters can potentially lead to poor discrimination.

The goal of the language model is to estimate prior probabilities of sentence models $P(M | \Theta)$. However, this language model is usually assumed to be independent of the acoustic models and is described in terms of an independent set of parameters Θ^* . At training time, Θ^* is learned separately, which is sub-optimal but convenient. The notion of “priors” (as defined above) however depends on how M is defined. For example, in continuous speech recognition, M usually represents a sequence of word models for which the probability $P(M | \Theta^*)$ can be estimated from large text corpora or from a given finite state automaton from which N -grams (i.e., the probability of a word given the $(N - 1)$ preceding words) are extracted. However, each word model is also represented in terms of an HMM that combines phone models according to the allowed pronunciations of that word; these multiple pronunciations can be learned from the data, from phonological rules, or from both. Likewise, each phone is also represented by a HMM that combines classes ω_c of Ω and for which the topology is usually chosen a priori independently of the data (or, sometimes, in a very limited way, e.g., to reflect minimum or average durations of the phones). Therefore, the grammar, the lexicon, and the phone models together comprise a kind of language model, specifying prior probabilities for sentences [$P(M)$], words, phones and HMM states [$P(q^k)$]. These priors are encoded in the topology and associated transition probabilities of the sentence, word and phone HMMs. Usually, it is preferable to infer word level priors from large text corpora, due to insufficient speech training material to derive so many parameters from the speech data. However, neural networks and discriminant training implicitly make use of these priors. As a consequence, if the priors observed on the training data are not the same as the priors that are given by the HMM topology (and which have been a priori given or trained from an independent knowledge source), there will be a mismatch that will impact recognition performance at the global level.

The discussion above should help to explain why it is necessary to divide the ANN outputs by the

class priors to get acceptable results with hybrid HMM/ANN systems (particularly when the test priors are different than the training priors, e.g., in the case of different words). However, it still does not explain why an embedded Viterbi training using local posteriors as such (i.e., without division by priors), for which the Viterbi convergence proof still holds, usually converges to a very poor segmentation, although there is no “prior mismatch” in this case (since ANN training data – i.e., segmentation – is assumed to have been obtained from a given and fixed underlying HMM topology). As independently reported by several laboratories (but not, to our knowledge, published) this kind of Viterbi training will indeed usually converge to a very poor local minimum corresponding to a completely unsatisfactory segmentation. In this case the class that had the initial highest prior will get all the nearby training vectors (up to the limit where one class gets all the frames, except the minimum number of frames associated with the other classes and as imposed by the topology of the model). Note that we observed this phenomenon in the framework of a Viterbi training, since at the time we did not know how to train a hybrid HMM/ANN system according to the “full” posterior (i.e., using all possible paths, and without performing explicit segmentation). This was another motivation of our REMAP work.

Finally, due to many causes, “standard” HMMs can also have “language leakage”. First, training is performed on some specific corpus, containing some specific word sequences that will inevitably bias the acoustic features used to train the acoustic models. It can also be shown that this leakage can actually get worse when using discriminant training (such as MMI) [79]. Furthermore, we note here that, after many assumptions, the acoustic score $p(X|M)$ is computed in terms of emission probabilities $p(x_n|q^k)$ and transition probabilities $P(q^l|q^k)$, where the latter certainly encode some kind of prior information at the state level (which implicitly contains information about the word sequences in the training data).

3.5.2. Merging acoustic and language models

In the framework of standard HMM and language models, we would like to challenge the reader a little bit further by questioning the use of Bayes’ rule (6)

in speech recognition. Bayes’ rule is actually used with two goals in mind:

1. To turn posteriors that are difficult to estimate into likelihoods that are (sometimes) easier to compute.
2. To have a generic way of splitting and combining acoustic and language information.

Of course this formulation is based on a particular model of speech recognition, that of classification over a noisy channel. However, given what has been discussed above, it is not clear to us what $P(M)$ and $p(X|M)$ really mean and which kind of information they encode. This seems to be confirmed by the fact that, to our knowledge, all current “state-of-the-art” systems are far from applying Bayes’ rule as such but that, because of many poorly understood factors (such as the underestimation of acoustic probabilities because of the assumption of temporal independence) empirical scaling factors have to be applied to $\log P(M)$ or $\log p(X|M)$. This is equivalent to raising the probabilities to a power, which is not particularly satisfying from the point of view of the statistical theory. In some sense, this simply means that the system does not work as we would expect it to work. This phenomenon is also related to the problem of the negligible impact that transition probabilities have in acoustic models. Of course, given the way they are usually computed, acoustic likelihoods are dimensional and dependent on the length of the sequence (resulting in very low values) while transition probabilities within and between words tend to have much larger values with a smaller range.

Of course, there are many ways to merge models or use different sources of information based on a (log) likelihood criterion (see, e.g., [20,43] to find some explanation why it is necessary to modify the logarithm of the joint likelihood of words and acoustics by weighting their contributions differently to reflect their respective accuracy) or a maximum entropy criterion [40].

Here we would like to ask the reader to consider a possible alternative to combine the acoustic match with language information, one which originates from a recent class of architectures referred to as a mixture of experts [61]. The general idea consists in considering that acoustic processing and grammatical analysis are performed by two independent

“experts”⁵, whose decisions are then combined in a later stage. More formally, this can be expressed as follows. Let us assume that we have two experts (i.e., models): an acoustic expert denoted E_x and a language expert denoted $E_{\mathcal{L}}$. Since these experts will be governed by independent set of parameters, we could also denote them Θ_x and $\Theta_{\mathcal{L}}$. Ideally, a speech recognition system should recognize the best model M associated with some acoustics X according to

$$P(M | X, \mathcal{L}), \quad (7)$$

in which \mathcal{L} represents what we know about the language model. In other words, we recognize a sentence in terms of what we hear (X) and what we know about the grammar (\mathcal{L}). Assuming that the two experts are mutually exclusive, i.e., $P(E_x) + P(E_{\mathcal{L}}) = 1$, in which $P(E_x) [P(E_{\mathcal{L}})]$ represents the probability that the acoustic [language] expert is more reliable than the language [acoustic] expert. In this case, (7) can be rewritten as

$$\begin{aligned} p(M | X, \mathcal{L}) &= P(M, E_x | X, \mathcal{L}) + P(M, E_{\mathcal{L}} | X, \mathcal{L}) \\ &= P(E_x | X, \mathcal{L}) P(M | E_x, X, \mathcal{L}) \\ &\quad + (1 - P(E_x | X, \mathcal{L})) P(M | E_{\mathcal{L}}, X, \mathcal{L}) \\ &\approx P(E_x | X, \mathcal{L}) P(M | E_x, X) \\ &\quad + (1 - P(E_x | X, \mathcal{L})) P(M | E_{\mathcal{L}}, \mathcal{L}), \end{aligned} \quad (8)$$

the last equation coming from the fact that it is assumed that the acoustic (linguistic) expert has no direct use of the language (acoustic) data. In (8), the first term represents the acoustic contribution and the second term the language contribution. The final score is thus obtained through a weighted sum of both (posterior) contributions, in which the weight $P(E_x | X, \mathcal{L})$ represents the reliability of the acoustic model given X and \mathcal{L} and could be a function of the signal-to-noise ratio or of some measure of discrepancy between the training and test data.

Of course this is a serious departure from standard

HMM systems since this would require computation of global posterior acoustic scores (containing the priors relative to information that is not in \mathcal{L} – see Section 3.3), and combination with the language score through a weighted sum. To our knowledge, this approach to merging acoustic and language models has not yet been explored.

4. Features for ASR

Most current speech recognizers derive their capabilities from training data. In the training stage the recognizer is presented with all the information that is available (i.e., both “signal” and “noise”). In principle, such exhaustive training should allow the separation of the desired signal from the noise during recognition.

However, the principle also calls for a model-free classifier (e.g., a nearest neighbor classifier) for which the need for data grows exponentially with the number of its free parameters. Since this is clearly rather impractical, even the most fervent believers in large training data sets do not present their classifier with raw speech time waveforms, but rather use some information-rate reduction prior to the classification, i.e., a feature extraction component (speech analysis).

Speech analysis in ASR is historically derived from speech coding applications (e.g., short-term Fourier transform, down-sampled filter bank output, LPC, ...). Both speech coding and ASR benefit from small feature sets, but otherwise needs in ASR and speech coding are often different. Generally speaking, speaker-dependent and environmental information should be preserved in speech coding to permit the high-quality re-synthesis of speech. This is generally not what is desired for ASR. In ASR, there is no need to reconstruct the original speech. Most often, the goal is to extract the linguistic message, and the nonlinguistic information sources introduce additional variability that needs to be dealt with either by training or during recognition (e.g., by the feature extraction). For most ASR tasks this unwelcome component would consist of the speaker characteristics, the environment in which the speech was produced, recording equipment used for speech acquisition, etc. Often, such sources of nonlinguistic

⁵ Or functional aspects of the brain. This also seems to make sense from a psychological point of view since, as we all know, it is often possible to perfectly recognize sentences even with extremely poor acoustics; on the other hand it is also possible to unambiguously recognize a nonsense sentence if the acoustics are good enough.

information are quite deterministic, and dealing with them by stochastic means in training appears to be unnecessary. They can sometimes be dealt with by a well designed speech analysis that would alleviate the variability resulting from non-linguistic information sources.

Thus, speech coding techniques may not be appropriate for ASR, and one should be free to design entirely different ways to extract useful features from the speech signal. Of course there is a catch here: information lost during feature extraction is lost for the recognition process and cannot be recaptured. As much as possible, we need to preserve information that can be used for decoding the linguistic message while maximally eliminating the remaining variability from the signal. In doing so, we also need to consider the properties of the subsequent stochastic model so that the feature extraction and the stochastic model match one another.

4.1. Auditory modeling

Human speech perception seems to be capable of focusing attention on the linguistic message during conversational speech. Is this primarily due to peripheral properties of human hearing, or is it a function of higher level processes? The answer is not entirely clear, but there is at least some evidence that several peripheral properties of human hearing might at least be partially responsible for the ways speech evolved and is being used for human communication [42].

Thus, if our goal is to eliminate components of the speech signal that are less relevant for the determination of linguistic component, a reasonable constraint to our design space would be to eliminate what human listeners cannot hear, while focusing attention on parts of the signal that are heard well. Thusfar, however, it appears that models of human hearing have not found good acceptance in ASR. We argue that there are at least three good reasons for this seeming failure:

- Failure to adapt the rest of the recognizer to new properties of auditory-like feature extraction.
- We should note, however, that in fact some of the most common feature sets used in ASR make use of at least some of the grossest properties of human hearing that researchers have found to be relevant to this work. Some of these will be discussed further on in this section.

4.1.1. How to increase error rate by auditory-like processing

New techniques are often tested on a well established task in a system which is finely tuned to some other technique. In a complex ASR system, there are many things that can go wrong, and usually at least one of them does when the new technique is substituted for an old one. Thus, e.g., the number of Gaussian mixtures and the number of states in an HMM model can be optimized for a particular feature extraction module. If this old module is replaced by a new one, the error rate might increase simply due to a mismatch between properties of the new feature extraction module and the remainder of the system, which has been tuned to the original features.

It is often better to run initial experiments on tasks that are as simple as possible (but – following the classic – not any simpler). Our early attempt for employing several basic properties of the human peripheral auditory system was Perceptual Linear Predictive (PLP) analysis [49]. An initial critical experiment that convinced us of the utility of PLP in speaker-independent ASR was a cross-speaker experiment in which the recognizer was trained on one speaker and tested on another [48]. PLP yielded about 50% accuracy and the competing technique (LPC) was grossly worse for the task. Several times we have been challenged for presenting these results “‘which did not relate to real ASR’”. To us, the experiment clearly showed that low-order PLP retains less of the speaker-specific characteristics than LPC (which has since been substantiated with many other experiments). Had we first tested this property on a system trained on many speakers to achieve the conventional 90 + % accuracy, (as we ultimately did), we probably would not have seen the characteristics of the new technique as clearly as we did from

- Testing on tasks that do not expose weaknesses of conventional feature extraction techniques.
- Ignoring the fact that not everything that hearing can do is necessarily used in human speech communication.

this 50% score for the relatively straightforward template-matching cross-speaker task.

4.1.2. *Should airplanes flap their wings?*

A common argument against the incorporation of knowledge about human hearing in ASR (or more generally the use of knowledge about sensory or cognitive processes to help in the design of pattern recognition machines) is: “Airplanes do not flap wings⁶; therefore automatic recognizers should not have ears!”. However, airplanes have wings and so do birds. Our understanding of the principles of aerodynamics that allow birds to stay in the air (e.g., Bernoulli’s Principle) also allows us to enjoy (or hate) the benefits of air travel. On the other hand, mimicry of nature without understanding the principle behind its design can cause us to fail.

Over the years, speech researchers have (knowingly or unknowingly) used certain properties of human hearing in design of speech engineering systems. We discuss several of these properties below.

- Nonlinear frequency warping – Decreasing selectivity of human hearing with frequency is one of the best documented and least disputed properties of human auditory perception. Bridle and Brown [18] and later Mermelstein [72], and Davis and Mermelstein [25] proposed the use of the cosine transform of logarithmic energies (cepstrum) from non-uniformly spaced bandpass filters with bandwidth increasing with frequency. Davis and Mermelstein proposed triangular filters with a shape which is roughly constant on the mel scale. The mel cepstrum is currently the dominant feature extraction technique in automatic speech recognition. Another practical technique that uses a similar Bark-warped frequency scale is Perceptual Linear Predictive (PLP) analysis [49].
- Root Spectral Compression – Perception of intensity appears to be consistent with a compressive type of nonlinearity. In particular, perceived loud-

ness of steady sound use is approximately proportional to the cubic root of its power [97].

Lim [68] investigated the use of different compressing functions in homomorphic analysis of speech. He concluded that cubic root compression was optimal with respect to the speech quality of re-synthesized speech. Hermansky et al. [51] experimented with varying compressive functions in linear predictive analysis and found that when the short-term power spectrum of speech is compressed using the cubic root function, the analysis is least affected by the fine spectral structure of voiced speech. Root spectral compression seems to help in modeling spectral envelope zeros, which occur in nasalized and fricative speech sounds. PLP analysis uses autoregressive modeling in the cubic-root compressed power spectral domain to derive a smooth approximation to the underlying frequency-warped auditory-like short-term spectrum. Furthermore, the root compressed power spectrum (root compression with exponents 2–4) appears to be optimal for processing that alleviates the effects of additive noise in the signal (see, e.g., [47,53,85]).

- Nonuniform spectral sensitivity of hearing – For typical levels of human speech communication, hearing is most sensitive in about the 1–3 kHz range, therefore emphasizing the second and third formant region.

To obtain more stable formant estimates, Itahashi and Yokoyama [56] proposed weighting the mel-warped speech spectral envelope by an approximation of the Fletcher–Munson equal loudness curve. Equal loudness weighting of the spectra was also found useful in PLP analysis. One can even argue that in conventional speech analyses, the typical 6 dB/oct preemphasis approximates the equal-loudness curve.

- Broad spectral integration – Klatt [65] speculated that for gender normalization, larger than 1 Bark spectral resolution would be required. This notion is supported by some perceptual studies that suggest that human speech perception could integrate formant peaks within 3.5 Bark interval [24], and therefore could merge several speech formants. Thus, frequency resolution for the perception of speech signals seems to be considerably broader than the critical-band concept would suggest.

⁶ According to Richard Lippmann, “After more than 10 years of analysis and model building, Prof. James DeLaurier at the University of Toronto has built a model airplane that flies by flapping its wings.” Perhaps we will now need a new metaphor ...

Pols et al. [81] reported that the first three (six) principal components of a set of non-uniformly spaced $1/3$ octave filter bank output power explain 82% (97%) of variance in his data. In later work Pols [82] also showed that these first three principal components can be used successfully in automatic speech recognition. The works of Pols et al. were apparently the inspiration for introducing the smoothing of mel-warped spectra by truncation of cosine series in mel cepstral analysis [19]. PLP analysis, in many other respects quite similar to mel cepstrum, does the final spectral smoothing by fitting it with the spectrum of an all-pole model, thus emphasizing major peaks in the auditory spectrum. Spectral smoothing has typically been found to be beneficial for ASR, especially with a limited amount of training data.

- Temporal properties of human hearing – All engineering models of peripheral phenomena discussed above operate on a single short-term (about 10–20 ms) window of speech. It is easy to show that this is not how human hearing works. Some researchers in auditory modeling realized this fact many years ago and began to explicitly model temporal properties of human hearing [21,94]. The main emphasis in these models seemed to be relatively short-term temporal phenomena. Increasingly, longer time-span phenomena are also being explored, as in some of the recent techniques discussed below.

4.2. Beyond 20 ms

It appears that the short-term memory of the auditory periphery in mammals (exhibited, e.g., by forward masking (see, e.g., [104]), the firing rate adaptation constant (see, e.g., [2]), and the buildup of loudness (see, e.g., [98])) is at least of the order of about 200 ms. This means that the peripheral human auditory system can effectively integrate rather large (about syllable sized) time-spans of the audio signal. This is an order of magnitude longer time-span than the span of temporal window of typical short-term speech analysis for ASR.

Before expanding this further, let us state three postulates:

1. Steady configurations of the vocal tract are rare and carry relatively little linguistic information.

2. Fundamental linguistic units are likely to be longer than 10 ms and one frame of short-term analysis result provides a description of its relatively short (quasi-stationary) part.

3. The communication channel and its noise level most often remains fixed or varies only rather slowly during the conversation.

If we accept these, we are ready to discuss some of weaknesses of the current recognition paradigm.

In standard HMM-based speech recognition, the speech signal is assumed to be a piecewise stationary sequence of short-term acoustic vectors. Usually these acoustic vectors are short-term spectral based vectors representing between 10 to 20 ms of the speech signal. Such a description of the speech signal as a succession of equally spaced short-term samples also originated in speech coding. It assumes that short-term segments of speech are independent samples from different and unrelated stationary processes. Since only a short-term “snapshot” of the signal is available at any given time, it is hard to distinguish between the “short-term quasi-stationary” signals (such as speech) and “long-term quasi-stationary” disturbances (such as fixed frequency characteristics of the communication channel or noise). The sequences of “short term, quasi-stationary” acoustic vectors are clearly correlated over time. According to the standard HMM formalism, the correlation between acoustic vectors observed on the same HMM state is not taken into account, which means that the general dynamics of the signal (inside each HMM state), including the time ordering of the acoustic vectors, is not utilized. Only larger time-span correlation (between subsequences of acoustic vectors) is modeled via the HMM topology, i.e., by the possible HMM state sequences allowed by the topology.

A classifier that does not utilize models of short-term temporal information could benefit from feature extraction that would use it. This potential for utilizing the medium-scale spectral dynamics is one of the major advantages of segment-based ASR. However, recently we have witnessed techniques for post-processing of short-term speech feature vectors which use local correlations between short-term feature vectors without any explicit segmentation. Several of these techniques that can be used with advantage in current frame-based systems are discussed below.

4.2.1. Multi-vector input

Makino et al. [71] proposed a so called time-spectrum pattern as a method to classify a short-term speech segment using additional information available in a longer time interval of speech surrounding a given speech instant. In this formulation, several short-term analysis vectors (spanning about 100–150 ms) are concatenated into one longer vector which is then treated as a basic input feature vector for the subsequent MLP classification. In [84] the dynamics was incorporated in the observations by appending to the feature vector observed at time n , the vector observed at time $n - \delta$ for some fixed offset δ . On several experiments, it was shown that this significantly increased recognition rates and that best results were obtained with a gap δ of 75 to 90 msec. Variants of these basic ideas are currently being used in some form by several ASR groups (see, e.g., [15,57,102]).

In hybrid HMM/ANN systems, it has been often shown that the use of multiple frames at the input of the network to compute local probabilities for HMM significantly improved both frame classification as well as word recognition performance. In [15] and the related original publications, systematic investigation of the “optimal” width of contextual input showed that time-spans between 30 and 90 ms generally led to the best recognition performance. However, we note here that this “optimal” value also depends on the number of phonemic HMM states; in [15] single state phonemic HMMs were used and 9 frames of 10 ms was chosen as the best contextual size while 3 state HMMs usually lead to a smaller contextual size.

Several works [20,45] reported successful use of relatively short time-spectral pattern with Linear Discriminant Analysis (LDA) [55]. Typically, in these cases a few adjacent frames are considered (typically 3–5 frames in total, spanning between 30 and 75 ms), on which LDA is performed to reduce the dimensionality of the acoustic features while minimizing the intra-class variance and maximizing the inter-class variance.

The multi-vector input technique relies on the classifier to discover the relative importance of time-advanced and time-delayed speech analysis vectors for the classification of the given speech instant. This has two advantages. Firstly, the speech models

can capture some of the correlation across adjacent frames. Secondly, in the presence of a disturbance, the classifier could, in principle, discover a fixed bias in the time trajectory of a given input feature, and discount it during recognition. The weights that are trained and applied to different time-shifted features form a multi-input multi-output (and in general nonlinear) filter applied to temporal trajectories of speech features.

The approaches discussed in this section all point in the direction of using more temporal and dynamic information in the recognition process.

4.2.2. Dynamic (*delta*) features

Furui [36] introduced dynamic features of speech to describe time trajectories of speech parameters in the vicinity of a given speech vector. He proposed the first three coefficients of the orthogonal polynomial representation of trajectories of cepstral coefficients within a certain time segment to describe (a) the mean value (later substituted by the cepstral vector in the center of the segment), (b) the slope, and (c) the curvature of trajectories within the given segment of cepstral feature representation.

Higher order dynamic features are invariant to any constant bias within the temporal window used for their derivation. Consequently, since the dynamic features are typically used on the cepstral coefficients, they are invariant to slowly varying linear (convolutive) distortions of the signal introduced, e.g., by different frequency characteristics of the communication environment.

The dynamic feature calculation represents a finite impulse response (FIR) filtering of the time trajectories of cepstral coefficients. The implied band-pass filters are rather selective, emphasizing speech components with a certain rate of change [52]. The emphasis on a limited range of rate-of-change effectively deemphasize feature components with other rates of change which nevertheless also carry the important linguistic information. It has been observed [32,36] that dynamic features do not perform too well on their own, and are therefore typically used only in conjunction with the original short-term (static) features.

In his early work [36], Furui found that relatively long time spans of about 160 ms for deriving delta features were optimal. This compares to an 85 ms

optimal window of Elenius and Blomberg [32] and to 75 to 90 ms for Poritz [83]. Later Furui used much shorter windows of about 50 ms for deriving the dynamic features [37], but Applebaum and Hanson [6] again observed the advantage of rather long temporal windows (of the order of 200 ms) in isolated-word recognition of noisy Lombard speech.

Dynamic features are used by virtually all state-of-the-art ASR systems and contribute possibly the only widely accepted significant departure from the simple frame-by-frame short-term feature extraction in ASR. This success may be attributed to the fact that dynamic features contribute new information that was previously unavailable to the pattern classification component of an ASR system: the information about the surroundings of the current short-term segment.

The use of dynamic features could make ASR systems more sensitive to speaking rate. This latter effect has sometimes been observed in practical real-time ASR systems, but to our knowledge has not systematically been studied and reported for artificial databases.

4.2.3. RASTA processing

Similarly to dynamic feature calculation, RASTA processing [52] filters the time trajectories of speech features. However, it differs from dynamic (“delta”) feature calculation in two respects:

1. RASTA includes processing sandwiched between two general nonlinearities applied to the feature space (implying nonlinear filtering), and
2. RASTA typically uses a rather broad band-pass filter with a relatively flat pass-band, which should allow for un-attenuated preservation of most speech components in the feature representation. Recent work [80] indicates that RASTA processing can simulate forward temporal masking in human hearing.

Ideally, after the appropriate feature transformation, the disturbing components should combine linearly with the components which need to be preserved, so that the two can be separated by linear RASTA filtering. One useful domain appears to be the critical-band spectrum with its amplitude compressed by the $y = \log(\text{const} + x)$ warping function, where the constant is proportional to the noise energy in the signal. The rationale behind this nonlinearity is to

allow for an approximately linear compression of low-energy spectral components (which contain noise that is additive in the linear domain) and for an approximately logarithmic compression of high-energy spectral components (which contain speech signal with possible linear distortions additive in the logarithmic domain).

The current RASTA bandpass filter is a four-zero, single-pole ARMA filter with the pass-band between approximately 1–12 Hz and with rather steep slopes. As most commonly applied, the integration time constant of the filter is about 160 ms. This filter pass-band appears to have some relation to properties of human hearing in perception of FM signals [41]. The steep filter slopes, however, do not seem to be consistent with human auditory perception [80].

Although the RASTA technique has been shown to be quite effective on some specific problems (involving different transfer functions or additive noise), it is still clearly at an early stage of development. It also raises many new issues and shows our poor understanding of the ideal speech recognition process.

Since the integrating time constant of the RASTA filter spans about the length of a typical demisyllable or syllable, RASTA processing makes speech features more dependent on a preceding context. Speech transitions are typically enhanced by RASTA processing. This seems (in principle) to be compatible with some properties of the human auditory system, and presents no problem in whole-word ASR. However, this property needs to be carefully considered in current sub-word (e.g., phoneme) based ASR systems where it is in obvious contradiction with standard HMM assumptions of piecewise stationarity.

Thus, on one hand we have in RASTA a simple technique that reduces the sensitivity of recognizers to environmental conditions, but on the other hand may not be entirely suitable for the current ASR recognition paradigm. Such a situation may be suggesting a need for further developments of the analysis technique so that it would perform better in a given recognizer. However, it is also possible (as we happen to believe) that in comparing the current speech instant with some of its history (and thus enhancing the transitional component in speech) the RASTA technique is doing the right thing. This suggests a need for a revision of standard speech

recognition technology, in order to be more consistent with modeling approaches that focus on transitions. The SPAM technique (discussed later) is our initial attempt at a solution to this problem.

4.2.4. Cepstral mean subtraction

One of the simplest signal processing operations that uses global knowledge is blind deconvolution [99]. In the original formulation of the technique an averaged spectrum of a new signal is matched to an averaged spectrum of some reference. The current reincarnation of this technique, known under the name of cepstral mean subtraction (see, e.g., [5]) simply removes the means of all time trajectories of cepstral coefficients, thus setting the log spectrum mean to zero. Cepstral mean subtraction is currently quickly becoming a mainstream supplementary technique in speech analysis in ASR whenever a real-time operation is not required.

To minimize the inherent processing delay of this technique, the mean can be computed over a relatively short amount of future or past data, and windows as short as 50 ms have been used [45]. Mean removal with the mean computed over different lengths of data implies different FIR filtering of the cepstral coefficient time trajectories [52]. The shape of the window has an effect on the implied frequency response of the process and the commonly-used square window may not be optimal.

Rosenberg and his colleagues recently reported that long windows did not perform as well as shorter (about 165 ms) ones [91]. Such a short time constant is consistent with the time constant of typical RASTA filters [52].

In similar experiments [46], short-term cepstral mean subtraction is performed by recursively re-estimating the cepstral mean over some previous frames (making it very similar to RASTA processing), followed by LDA over 2 frames⁷. Here, it was also found that the optimal “memory” for the mean computation should be around 250 ms.

4.2.5. Dynamic cepstrum

As with dynamic feature calculation, a so called dynamic cepstrum [3] explicitly imposes weighting

of cepstral coefficients within the relatively short (about 50 ms) temporal window. The weighting represents a Gaussian approximation of the data on simultaneous time-frequency masking. It implies longer time constants for forward temporal masking on lower frequency cepstral coefficients. The dynamic cepstrum calculation resembles the dynamic feature calculation since it can be also interpreted as an FIR filtering of cepstral coefficient trajectories. Different cepstral coefficients are processed by different FIR filters. All the filters are high-pass filters with a cutoff frequency around 10–15 Hz. The filters for lower cepstral coefficients have steeper slopes than filters for higher cepstral coefficients.

4.2.6. Probabilistic optimum filtering

In this technique [77], relatively short (typically about 70 ms) two-dimensional piecewise-linear FIR filters are used on a time-frequency pattern of cepstral coefficients. The filters are derived by optimizing the mapping between a vector-quantized vector space of clean speech cepstral coefficients and a vector space of noisy coefficients. The final filter is a weighted combination of linear filters using an estimate of the current noise level to determine the weights. Use of two dimensional time-frequency filters allows for exploitation of inter-channel dependencies.

4.2.7. Discussion

Based on the evidence discussed in this section, we conclude that there is a case for looking beyond the standard 10 ms speech frames, and using larger time-spans of the speech signal (ranging between 30 and 200 ms) for recognition.

It may be possible that there are several reasons for getting to this general conclusion. The first phenomenon, at the level of 30 to 90 ms, could be related to a better modeling of the time correlation between successive acoustic vectors observed in each HMM state. Depending on the number of states used to represent a phoneme, the time spent (hence the correlation to be captured) on each state will span between 30 and 90 ms.

The second phenomenon takes place between 100 and 200 ms and could be associated with more global properties of human hearing such as short-term memory of auditory periphery, firing rate adaptation

⁷ See comments above on multi-vector inputs. However, the topology of the HMM being used in [46] is not described.

constant, or forward temporal masking effects (see, e.g., [50] for further discussion of this). The practical engineering argument for longer time constants (e.g., the 150–200 ms constants for both RASTA and cepstral mean subtraction) is the following: since the goal is separation of “stationary” disturbances from the (long-term) “non-stationary” speech signal, neither RASTA filtering nor cepstral mean subtraction should be applied on a speech segment that is so short that it could only include a single steady-state phoneme. Indeed, in this case, estimating and subtracting the cepstral mean on such a short segment would result in largely eliminating distinguishing characteristics of the whole segment. The processing thus needs to “see” a portion of the speech signal that is sufficiently large enough to evaluate the segment characteristics Relative (“R” as in RASTA) to some previous speech event, which is the case when spanning a part of the speech signal of the size of a demisyllable or syllable.

5. Beyond conventional HMM recognition

Assuming that we have piqued the reader’s interest in some alternatives to conventional HMM systems, we describe a few that we know of.

5.1. Modeling speech dynamics: stochastic segments, stochastic templates, non-stationary HMMs and autoregressive HMMs

One attempt for addressing the inconsistency of the piecewise stationary assumption of conventional HMMs was a proposal of Ghitza and Sondhi [38] to use the diphone, represented by a stochastic description of its spectral dynamics as a basic unit in HMM recognition. Their approach so far requires a hand-segmented diphone training database.

Stochastic segment modeling [78] has also been explored as an alternative to the conventional HMM paradigm. For practical reasons, this method has thus far relied on segmentation obtained from standard HMM systems.

Deng [29] suggests a so-called non-stationary HMM, which represents a phoneme not only by its static properties but also by its first order (and sometimes higher order) dynamics.

Another explicit (but still stationary) model of speech dynamics, which in principle does not need any pre-segmentation, is the (multivariate) autoregressive HMM (AR-HMM). Initially defined on the basis of (HMM state dependent) linear functions [62,83], those models were recently generalized to non-linear AR models where the nonlinear function associated with each HMM state is implemented by a neural network (minimizing during training the square error between the predicted acoustic vector (based on the p previous vectors at the input of the net) and the actual acoustic vector at time n [67,101]. Logarithmic emission likelihoods are then estimated as prediction errors of (class-dependent) AR models (of order p). Indeed, for both linear and nonlinear AR models, it can be shown that if the prediction errors (driving noise of the AR models) are assumed to be Gaussian, independent, identically distributed random variables with zero mean⁸, prediction errors⁹ can be considered as log likelihoods that can be used in standard HMMs. Similarly, in [103], the correlation across several frames is explicitly modeled with a multivariate, full covariance matrix, Gaussian density defined over two consecutive acoustic vectors¹⁰.

Although AR-HMM techniques clearly have the potential advantage of modeling the dynamics (time correlation) of the speech signal (although still limited by stationarity assumptions), none of those models ever led (at least to our knowledge) to conclusive experimental results for reasons that have never been clearly identified (see, e.g., [26]). So far, we have not observed reports of these methods providing improvements over standard HMM or hybrid HMM/ANN approaches. Some plausible explanations for this discrepancy between theory and practical results include:

1. Increase in the number of parameters.
2. Estimating autoregressive models implicitly assumes some “smoothness” properties of the sig-

⁸ And, sometimes, a unity covariance matrix, although this is not really necessary.

⁹ Or weighted prediction errors in the case of non-unity covariance matrix.

¹⁰ This can be shown equivalent to estimating a multivariate autoregressive process [103].

nal, which is not always true in the case of speech (and, consequently, what is gained on the one hand is lost on the other).

Finally, we note an interesting convergence that has taken place between research systems. Ten years ago there was a seemingly unbridgeable gap between systems that were based on acoustic-phonetic knowledge (making use of phonetic segments) [105] and frame-based statistical systems. Since then, there has been an increased (though still minority) interest in statistical techniques such as those mentioned above, and this has been aided by the ability of the frame-based techniques to generate N -best lists and lattices that can summarize plausible segmentations. In the meanwhile, segment-based systems such as those described in [44,57] now routinely make use of statistical methods such as hybrid HMM/ANNs or likelihood-based sequence models. Perhaps the incorporation of segmental and suprasegmental features can be explored more effectively now that we have powerful statistical mechanisms in place that might show us how to incorporate this kind of knowledge.

5.2. Stochastic perceptual auditory model (SPAM)

It is known that in human speech recognition, the perceptually-dominant and information-rich portions of the speech signal, which may also be the parts with a better chance to withstand adverse acoustical conditions, are the (phonetic) transitions (see, e.g., [36] for some experimental evidence).

As noted above, a first step in this direction was to use highpass or bandpass filtering of critical band trajectories (RASTA processing) to emphasize transitions [52]. While this can be helpful in reducing errors due to (channel) mismatches between training and testing conditions, the resulting observation sequence is a representation that has emphasized the regions of strong change and de-emphasized temporal regions without significant spectral change. This is a mismatch to the underlying speech model in standard HMMs, which has been designed to represent piecewise stationary signals. In general, modeling transitions or any non-stationary properties of speech signal require major modifications of standard HMMs. Therefore, it is likely that transition-

based systems will require a fundamentally different kind of underlying statistical model.

To address the above problem we proposed to model speech as a succession of auditory events or “avents”, separated by relatively stationary periods (ca. 50–150 ms). Avents correspond to times when the spectrum and amplitude are rapidly changing, which are believed to be the most important regions for phonetic discrimination [36]. The stationary periods are mapped to a single tied state, and so modeling power is focused on regions of significant change [75]. This approach is referred to as Stochastic Perceptual Auditory-event-based Model (SPAM).

In this case, SPAMs are defined from a set of avent classes $\Omega = \{\omega_0, \omega_1, \dots, \omega_c\}$, in which all ω_c 's, for $c \neq 0$, represent avent classes and ω_0 represents the non-avent class or non-perceiving state. This set is currently initialized to correspond to truncated diphones; that is, phone boundaries with the local region of the time series associated with them. Given such an initialization, the avents would be determined automatically in an embedded Viterbi-based dynamic programming procedure (as is currently accomplished with phone-like subword models).

As for HMMs, each word to be recognized is thus represented by a SPAM model defined as an oriented graph M with let say K avent states q^k , each of them being associated with an avent class $\omega(q^k) \in \Omega$. As initially discussed in [75], and more recently in [76], and starting from (1)–(3) [i.e., the actual speech recognition model, as opposed to production models based on $p(X|M)$], one can do SPAM recognition based on the following local acoustic probabilities:

$$P[\omega(q^l) | \omega(q^k), \Delta(n), X_{n-c}^{n+d}] \quad (9)$$

$\forall l = 0, 1, \dots, K$ and $\forall k = 1, 2, \dots, K$, and in which $X_{n-c}^{n+d} = \{x_{n-c}, \dots, x_n, \dots, x_{n+d}\}$ represents the current chunk of input data and $n - \Delta n$ the previous time index for which an avent had been hypothesized and becomes one of the stochastic variables of the model. During training, these local probabilities are estimated by a neural network, via an iterative Viterbi-like segmentation to provide the net with targets or via REMAP (as explained in Section 3.3). According to our SPAM constraints, these local probabilities are used for training and decoding in

particular left-to-right HMMs constituted by sequences of *av* (diphone) states (with no loop allowed) separated by (looped) tied non-perceiving states.

Preliminary experiments on isolated digits were reported in [76]. In these experiments, it was shown that our best phone-based system had about half of the error rate of SPAM for the clean digits. As of this writing, our SPAM-based systems are now as good as our phone-based ones for clean digits as well.

However, for the noisy case (10 dB SNR), the performance of the two systems were comparable. Combination of both approaches led to similar performance for the clean case but to a 30% improvement for the noisy case, which shows again that it is sometimes worth pursuing “deviant” recognition paradigms. On top of this one may have the pleasant feeling to have learned something.

We note here that SPAM is somewhat related to the variable frame-rate analysis method used since the late seventies [100] in which, when the acoustic front-end produces a sequence of similar consecutive frames, only the first frame together with the length of the sequence is passed to the training and recognition process. Thus, this also emphasizes dynamic portions of the speech signal and remove correlation between successive frames. In [92], this was shown to reduce error rates. A related approach [32] explicitly emphasized spectral transitions and reported a slight improvement in error rate. The goal of SPAM is similar except that in this latter case the HMM model has been modified to take only transitions into account, with only one transition associated with each perceptual state while stationary sequences are matched on the same (non-perceiving) state.

5.3. Multi-band ASR

The work of Fletcher and his colleagues [35] (see the insightful review of his work in [4]) suggests that the decoding of the linguistic message is based on decisions within narrow frequency bands that are processed quite independently of each other. Recombination of decisions from those frequency bands is done at “certain levels” so that the global error rate is equal to the product of “band-limited” error rates within the independent frequency channels. This also

means that if any of the frequency bands yield zero (or low) error rate, the resulting error rate would be also zero (or relatively low), (almost) independently of the error rates in the remaining bands.

We see at least three engineering reasons for considering this proposal:

1. Different strategies for decoding the linguistic message may be used in different frequency bands.
2. Additive noise may contaminate only some specific frequency bands. When the recognition is based on several independent decisions from different frequency bands, the decoding of linguistic message does not have to be severely impaired, as long as the remaining clean bands supply sufficiently reliable information.
3. Temporal asynchronies between different channels (including ones that may be derived from variability in the acoustic environment) may potentially be handled better by a system that is extracting higher level information before the channels are recombined.

There are several practical problems associated with this proposition, the most urgent ones being:

- It is not clear how to subdivide the available input data.
- It appears that such a technique would assume some knowledge about the reliability of each frequency band.
- It is not clear at which decision level (microsegment, phoneme, diphone, syllable, word, sentence, message, ...) the recombination should take place.
- It may be best for the different frequency bands to be processed differently, and recombined at a later stage.

Preliminary work using (or attempting to simulate) some of these properties have recently been reported. For example, in [31], tentative decisions (or labels) of the identity of a given speech element were assigned in parallel by sub-recognizers, each operating on a band-limited portion of the speech waveform¹¹. The outputs of these independent channels were subsequently combined to render the final decision. Although this approach did not lead to real

¹¹ In this case, 4 independent frequency bands were used.

improvement, it was quite remarkable that it did not increase error rates, although the recombination scheme was quite simple and no optimization of the frequency bands had been performed. Also, the resulting system was not tested under noisy conditions (under which this kind of approach should prove to be most interesting).

So far, initial experiments with multi-band ASR in our laboratories have yielded two tentative conclusions:

- Outputs from two recognizers working on two (slightly overlapping) frequency bands can be combined in such a way that the resulting multi-band recognizer appears to perform noticeably better than a single broad-band recognizer.
- When one of the frequency bands is contaminated by selective noise, the optimal deemphasis of the output of the contaminated recognizer in the multiband combination yields much more graceful degradation than the broad-band recognizer alone.

For further discussion on this topic and preliminary results, we refer the reader to [106].

6. Knowing when we do not know

Finally, one certain way to increase the error rate¹² (while actually improving system utility) is to reject responses that fail to satisfy certain criteria. ASR research has often focused on controlled tasks in which test data have few or no out-of-vocabulary words or sounds. It may have been beneficial to our work if the original targeted ASR research task was not recognition of several digits but rather word-spotting for a single word (recognizing a word in the universe of all other sounds). Just as we evaluate speaker recognition in the presence of impostors, we also need to be evaluating ASR technology for its rejection capability in the presence of out-of-vocabulary sounds. Some work on this important problem has been initiated [7] but to our knowledge this problem still does not get sufficient attention.

Error rate is, of course, a useful measure for

speech recognition researchers. It is straightforward to compute, and in general is well correlated with other performance measures. It is often the most straightforward way to find our path in a research that is not yet well understood. However, it has its limitations. While it is difficult to find another simple measure that can be so straightforwardly applied to artificial experiments, speech scientists and engineers should note that at least one significant purpose for their work is to develop the technology that enables speech recognition systems to work on real applications. For the users of ASR technology, the most significant measure is really the success of task completion, and this can often be aided more effectively by our knowing when we don't know, rather than just making our errors less frequent [13]. In some applications it may be preferable to achieve 85% accuracy with high confidence than to achieve 99% accuracy in which we never know whether the result is correct or not. While practical ASR systems generally use some measure of confidence, the problem of reliable confidence measures receives very little attention in the research community. It is likely that as the focus of our research shifts towards attacking realistic concrete tasks it would also raise the profile of some of these issues.

7. Conclusions

Throughout this paper, our discussion was (unavoidably) biased toward the systems we have studied, and we have undoubtedly neglected many other innovative approaches to ASR. We do hope that our colleagues whose work we neglected to mention will forgive us, but that their voices will be heard in discussions that follow. After all, the need for discussion is our main reason for writing this paper.

Here we have primarily intended to challenge the reader to consider more radical deviations from current wisdom in order to improve our overall understanding. Some motivations for the specific approaches we have described in this paper have come from statistical theory, properties of human hearing, and other sources of inspiration. A common factor in much of our current directions is the attempt to make use of longer time spans of the speech signal, both in the statistical methods and in the feature extraction.

¹² Here we mean the error rate defined as number of correct responses relative to number of trials.

Other concerns include a desire to better represent the relationships between external language models and the implicit language information represented in the acoustic models. We encourage the reader to incorporate both the specific research directions we have described here and others that we have not mentioned (and/or are not aware of) to branch out and better explore the space of possible solutions to problems in ASR. In all of these cases, there is much more unknown than known, and so each new attempt to deal with these issues is likely to result in a higher error rate for initial tests. We hope that the reader (and those who might fund the reader¹³!) will have the patience to wait through such a period in order to provide deeper understanding of the issues that good experiments can bring.

The current level of difficulty for mastery of the larger systems is so high that some in the field have tended to discourage newcomers. However, we believe that ASR is still a field in which many of the basic problems are still barely touched, and in which new ideas from newcomers will be critical. We should welcome these ideas, even if they increase the error rate.

We conclude by recalling an old story about Thomas Edison: After he (or some unsung hero in his lab) had invented the light bulb, a reporter asked him how many light bulbs he had built before he found one that worked.

“He replied, ‘About one thousand.’

The reporter then asked, ‘How did it feel to fail 999 times?’

Edison replied, ‘I didn’t fail 999 times; I was involved in a process that required 1,000 steps.’”

(recalled by Norris Parker Smith, a journalist who specializes in HPC and high-bandwidth communications.)

Acknowledgements

This work has benefited from collaboration and discussions with many of our friends and colleagues,

but particularly from our interactions with Steven Greenberg of UCB and ICSI. We also thank the many organizations that have provided support for our work over the years, and by doing so demonstrated their patience while we explored many techniques that successfully increased our error rates. While not complete, this list includes: the Joint Services Electronics Program (JSEP), through contract F49620-94-C-0038 to UC Berkeley; the National Science Foundation and ARPA, through IRA-9314959; the European Union, through basic research grants to projects Wernicke (6487) and SPRACH (2077); and primary support from our institutions.

Finally, we would like give thanks to our reviewers, whose comments have pushed us to (in our opinion) greatly improve this paper; and to our Editor, Christel Sorin, who has been very helpful throughout this process.

References

- [1] S.M. Ahadi and P.C. Woodland (1995), “Rapid speaker adaptation using model prediction”, *Proc. IEEE Internat. Conf. Acoust. Speech, Detroit, MI*, pp. 684–687.
- [2] L. Aitkin, C. Dunlop and W. Webster (1966), “Click-evoked response patterns of single units in the medial geniculate body of the cat”, *J. Neurophysiology*, Vol. 29, pp. 109–123.
- [3] K. Aikawa, H. Singer, H. Kawahara and Y. Tohkura (1993), “A dynamic cepstrum incorporating time-frequency masking and its application to continuous speech recognition”, *Proc. IEEE Internat. Conf. Acoust. Speech Signal Process., Minneapolis, MN*, pp. II-668–671.
- [4] J.B. Allen (1994), “How do humans process and recognize speech?”, *IEEE Trans. Speech Audio Process.*, Vol. 2, No. 4, pp. 567–577.
- [5] A. Anastasakos, F. Kubala, J. Makhoul and R. Schwartz (1994), “Adaptation to new microphones using tied-mixture normalization”, *Proc. IEEE Internat. Conf. Acoust. Speech Signal Process., Adelaide, Australia*, pp. I-433–437.
- [6] T.H. Applebaum and B.A. Hanson (1991), “Regression features for recognition of speech in quiet and in noise”, *Proc. IEEE Internat. Conf. Acoust. Speech Signal Process., Toronto, Canada*, pp. 985–989.
- [7] A. Asadi, R. Schwartz and J. Makhoul (1990), “Automatic detection of new words in a large vocabulary continuous speech recognition system”, *Proc. IEEE Internat. Conf. Acoust. Speech Signal Process., Albuquerque, NM*, pp. 125–128.
- [8] S. Austin, G. Zavaliagos, J. Makhoul and J. Schwartz (1992), “Improving state-of-the-art continuous speech

¹³ Not to mention the authors ...

- recognition systems using the N -best paradigm with neural networks”, *Proc. DARPA Speech and Natural Language Workshop, Harriman, NY* (Morgan Kaufmann, Los Altos, CA), pp. 180–184.
- [9] L.R. Bahl, P.F. Brown, P.V. de Souza and R.L. Mercer (1988), “A new algorithm for the estimation of hidden Markov models”, *Proc. IEEE Internat. Conf. Acoust. Speech Signal Process., New York, NY*, pp. 493–496.
- [10] J.K. Baker (1975), “The Dragon system – An overview”, *IEEE Trans. Acoust. Speech Signal Process.*, Vol. ASSP-23, No. 1, pp. 24–29.
- [11] Y.R. Bengio, R. De Mori, G. Flammia and R. Kompe (1992), “Global optimization of a neural-hidden Markov model hybrid”, *IEEE Trans. Neural Networks*, Vol. 3, pp. 252–258.
- [12] Y. Bengio and P. Frasconi (1995), “An input output HMM architecture”, in *Advances in Neural Information Processing Systems*, ed. by G. Tesauro, D. Touretzky and T. Leen, Vol. 7 (MIT Press, Cambridge, MA).
- [13] A. Bounds (1995), Personal communication.
- [14] H. Boulard and C.J. Wellekens (1990), “Links between Markov models and multilayer perceptrons”, *IEEE Trans. Pattern Anal. Machine Intell.*, Vol. 12, No. 12, pp. 1167–1178.
- [15] H. Boulard and N. Morgan (1994), *Connectionist Speech Recognition – A Hybrid Approach* (Kluwer Academic Publishers, Dordrecht).
- [16] H. Boulard, Y. Konig and N. Morgan (1994), REMAP: Recursive estimation and maximization of a posteriori probabilities – Application to transition-based connectionist speech recognition, ICSI Technical Report TR94-064, Internat. Computer Science Institute, CA.
- [17] H. Boulard, Y. Konig and N. Morgan (1995), “REMAP: recursive estimation and maximization of a posteriori probabilities in connectionist speech recognition”, *Proc. Eurospeech '95, Madrid, Spain*.
- [18] J.S. Bridle and M.D. Brown (1974), An experimental automatic word recognition system, JSRU Report No. 1003, Ruislip, England: Joint Speech Research Unit.
- [19] J.S. Bridle (1995), Personal communication.
- [20] P. Brown, The acoustic-modelling problem in automatic speech recognition, PhD Thesis, Computer Science Department, Carnegie Mellon University.
- [21] J.R. Cohen (1989), “Application of an auditory model to speech recognition”, *J. Acoust. Soc. Amer.*, Vol. 85, No. 6, pp. 2623–2629.
- [22] J.R. Cohen (1995), Informal communication.
- [23] M. Cohen, H. Franco, N. Morgan, D. Rumelhart and V. Abrash (1992), “Hybrid neural network/hidden Markov model continuous speech recognition”, *Proc. Internat. Conf. Speech Language Processing, Banff, Canada*, pp. 915–918.
- [24] L.A. Chistovich (1985), “Central auditory processing of peripheral vowel spectra”, *J. Acoust. Soc. Amer.*, Vol. 77, pp. 789–805.
- [25] S.B. Davis and P. Mermelstein (1980), “Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences”, *IEEE Trans. Acoust. Speech Signal Process.*, Vol. 28, No. 4, pp. 357–366.
- [26] P. de La Noue, S. Levinson and M. Sondhi (1989), Incorporating the time correlation between successive observations in an acoustic-phonetic hidden Markov model for continuous speech recognition, AT&T Technical Memorandum No. 11226.
- [27] S. Della Pietra, V. Della Pietra, R.L. Mercer and S. Roukos (1992), “Adaptive language modelling using minimum discriminant estimation”, *Proc. DARPA Speech and Natural Language Workshop, Harriman, NY*, pp. 103–106.
- [28] A.P. Dempster, N.M. Laird and D.B. Rubin (1977), “Maximum likelihood from incomplete data via the EM algorithm”, *J. Roy. Statist. Soc.*, Vol. 39, pp. 1–38.
- [29] L. Deng, M. Aksmanovic, X. Sun and C. Wu (1994), “Speech recognition using hidden Markov models with polynomial regression functions as nonstationary states”, *IEEE Trans. Speech Audio Process.*, Vol. 2, No. 4, pp. 507–520.
- [30] G. Doddington (1989), “Phonetically sensitive discriminants for improved speech recognition”, *Proc. IEEE Internat. Conf. Acoust. Speech Signal Process., Glasgow, Scotland*, pp. 556–559.
- [31] P. Duchnowski (1993), A new structure for automatic speech recognition, PhD Thesis, MIT.
- [32] K. Elenius and M. Blomberg (1982), “Effects of emphasizing transitional or stationary parts of the speech signal in a discrete utterance recognition system”, *Proc. IEEE Internat. Conf. Acoust. Speech Signal Process., Paris, France*, pp. 535–537.
- [33] M. Fany, P. Schmid and R. Cole (1993), “City name recognition over the telephone”, *Proc. IEEE Internat. Conf. Acoust. Speech Signal Process., Minneapolis, MN*, pp. 1-549–552.
- [34] M.A. Franzini, K.F. Lee and A. Waibel (1990), “Connectionist Viterbi training: A new hybrid method for continuous speech recognition”, *IEEE Proc. Internat. Conf. Acoust. Speech Signal Process., Albuquerque, NM*, pp. 425–428.
- [35] H. Fletcher (1953), *Speech and Hearing in Communication* (Krieger, New York).
- [36] S. Furui (1981), “Cepstral analysis technique for automatic speaker verification”, *IEEE Trans. Acoust. Speech Signal Process.*, Vol. 29, pp. 254–272.
- [37] S. Furui (1986), “Speaker independent isolated word recognizer using dynamic features of speech spectrum”, *IEEE Trans. Acoust. Speech Signal Process.*, Vol. 34, No. 1, pp. 52–59.
- [38] O. Ghitza and M.M. Sondhi (1993), “Hidden Markov models with templates as non-stationary states: An application to speech recognition”, *Computer Speech and Language*, Vol. 2, pp. 101–119.
- [39] H. Gish (1990), “A probabilistic approach to the understanding and training of neural network classifiers”, *Proc. IEEE Internat. Conf. Acoust. Speech Signal Process., Albuquerque, NM*, pp. 1361–1364.

- [40] I.J. Good (1963), “Maximum entropy for hypothesis formulation, especially for multidimensional contingency tables”, *Ann. Math. Statist.*, Vol. 34, pp. 911–934.
- [41] G.G.R. Green (1976), Temporal aspects of audition, PhD Thesis, Oxford, UK.
- [42] S. Greenberg (1988), “The representation of speech in the auditory periphery”, *J. Phonetics*, Vol. 16, pp. 1–151.
- [43] P.S. Gopalakrishnan, D. Kanecsky, A. Nada, D. Nahamoo and M.A. Picheny (1988), “Decoder selection based on cross-entropies”, *Proc. IEEE Internat. Conf. Acoust. Speech Signal Process.*, New York, NY, pp. 20–23.
- [44] W.D. Goldenthal (1994), Statistical trajectory models for phonetic recognition, PhD Thesis, MIT.
- [45] R. Haeb-Umbach, D. Geller and H. Ney (1994), “Improvements in connected digit recognition using linear discriminant analysis and mixture densities”, *Proc. IEEE Internat. Conf. Acoust. Speech Signal Process.*, Adelaide, Australia, pp. II-239–242.
- [46] A. Hauenstein and E. Marschall (1995), “Methods for improved speech recognition over telephone line”, *Proc. IEEE Internat. Conf. Acoust. Speech Signal Process.*, Detroit, MI, pp. 425–428.
- [47] B. Hanson and D. Wong (1984), “The harmonic magnitude suppression (HMS) technique for intelligibility enhancement in the presence in interfering speech”, *Proc. IEEE Internat. Conf. Acoust. Speech Signal Process.*, pp. 18.A.5.1–18.A.5.4.
- [48] H. Hermansky (1987), “An efficient speaker-independent automatic speech recognition by simulation of some properties of human auditory perception”, *Proc. IEEE Internat. Conf. Acoust. Speech Signal Process.*, Dallas, TX, pp. 1159–1162.
- [49] H. Hermansky (1990), “Perceptual linear predictive (PLP) analysis of speech”, *J. Acoust. Soc. Amer.*, Vol. 87, No. 4, pp. 1738–1752.
- [50] H. Hermansky (1995), “Exploring temporal domain for robustness in speech recognition”, *Proc. 15th Internat. Congress on Acoustics, Trondheim, Norway*, Vol. II., pp. 61–64.
- [51] H. Hermansky, H. Fujisaki and Y. Sato (1983), “Analysis and synthesis of speech based on spectral transform linear predictive method”, *Proc. IEEE Internat. Conf. Acoust. Speech Signal Process.*, Boston, MA, pp. 777–780.
- [52] H. Hermansky and N. Morgan (1994), “RASTA processing of speech”, *IEEE Trans. Speech Audio Process.*, Vol. 2, No. 4, pp. 578–589.
- [53] H. Hermansky, E. Wan and C. Avendano (1995), “Speech enhancement based on temporal processing”, *Proc. IEEE Internat. Conf. Acoust. Speech Signal Process.*, Detroit, MI, pp. 405–408.
- [54] M.M. Hochberg, S.J. Renals, A.J. Robinson and G.D. Cook (1995), “Recent improvements to the ABBOT large vocabulary CSR system”, *Proc. IEEE Internat. Conf. Acoust. Speech Signal Process.*, Detroit, MI, pp. 69–72.
- [55] M. Hunt and C. Lefebvre (1989), “A comparison of several acoustic representations for speech recognition with degraded and undegraded speech”, *Internat. Conf. Acoust. Speech Signal Process.*, Glasgow, Scotland, pp. 262–265.
- [56] S. Itahashi and S. Yokoyama (1976), “Automatic formant extraction utilizing mel scale and equal loudness contour”, *Internat. Conf. Acoust. Speech Signal Process.*, Philadelphia, PA, pp. 310–313.
- [57] R.D.T. Janseen, M. Fauty and R.A. Cole (1991), “Speaker independent phonetic classification in continuous English letters”, *Proc. Internat. Joint Conf. on Neural Networks*, Seattle, WA, pp. II-801–808.
- [58] F. Jelinek, L.R. Bahl and R.L. Mercer (1975), “Design of a linguistic statistical decoder for the recognition of continuous speech”, *IEEE Trans. Information Theory*, Vol. IT-21, pp. 250–256.
- [59] F. Jelinek (1976), “Continuous speech recognition by statistical methods”, *IEEE Proc.*, Vol. 64, No. 4, pp. 532–556.
- [60] F. Jelinek, B. Meriardo, S. Roukos, and M. Strauss (1991), “A dynamic language model for speech recognition”, *Proc. DARPA Speech and Natural Language Workshop*, Pacific Grove, CA, pp. 293–295.
- [61] M.I. Jordan and R.A. Jacobs (1994), “Hierarchical mixtures of experts and the EM algorithm”, *Neural Computation*, Vol. 6, pp. 181–214.
- [62] B.H. Juang and L.R. Rabiner (1985), “Mixture autoregressive hidden Markov models for speech signals”, *IEEE Trans. Acoust. Speech Signal Process.*, Vol. 33, No. 6, pp. 1404–14013.
- [63] T. Kamm, A.G. Andreou and J. Cohen (1995), “Vocal tract normalization in speech recognition: Compensation for systematic speaker variability”, *Proc. 15th Annual Speech Research Symposium*, Johns Hopkins University, Baltimore, MI, pp. 175–179.
- [64] S. Katagiri, C.H. Lee and B.H. Juang (1991), “New discriminative training algorithms based on the generalized probabilistic descent method”, in *Proc. IEEE Workshop on Neural Networks for Signal Process.*, edited by B.H. Juang, S.Y. Kung and C.A. Kamm (Morgan Kaufman, Los Altos, CA), pp. 299–308.
- [65] D.H. Klatt (1982), “Speech processing strategies based on auditory models”, in *The Representation of Speech in the Peripheral Auditory System*, ed. by R. Carlson and B. Granstrom (Elsevier – Biomedical Press, New York), pp. 181–202.
- [66] C-H. Lee, C-H. Lin and B-H. Juang (1991), “A study on speaker adaptation of the parameters of continuous density hidden Markov models”, *IEEE Trans. Signal Process.*, Vol. 39, No. 4, pp. 806–814.
- [67] E. Levin (1993), “Hidden control neural architecture modeling of nonlinear time varying systems and its applications”, *IEEE Trans. Neural Networks*, Vol. 4, No. 1, pp. 109–116.
- [68] J.S. Lim (1979), “Spectral root homomorphic deconvolution system”, *IEEE Trans. Acoust. Speech Signal Process.*, Vol. 27, No. 3, pp. 223–233.

- [69] L.A. Liporace (1982), "Maximum likelihood estimation for multivariate observations of Markov sources", *IEEE Trans. Information Theory*, Vol. IT-28, No. 5, pp. 729–734.
- [70] D.M. Lubensky, A.O. Asadi and J.M. Naik (1994), "Connected digit recognition using connectionist probability estimators and mixture-Gaussian densities", *Proc. Internat. Conf. on Spoken Language Processing, Yokohama, Japan*.
- [71] S. Makino, T. Kawabata and K. Kido (1983), "Recognition of consonant based on the Perceptron model", *Proc. IEEE Internat. Conf. Acoust. Speech Signal Process., Boston, MA*, pp. 738–741.
- [72] P. Mermelstein (1976), "Distance measures for speech recognition, psychological and instrumental", in *Pattern Recognition and Artificial Intelligence*, ed. by R.C.H. Chen (Academic Press, New York), pp. 374–388.
- [73] N. Morgan and H. Boulard (1990), "Continuous speech recognition using multilayer perceptrons with hidden Markov models", *IEEE Proc. Internat. Conf. Acoust. Speech Signal Process., Albuquerque, NM*, pp. 413–416.
- [74] N. Morgan and H. Boulard (1995), "Neural networks for statistical recognition of continuous speech", *Proc. IEEE*, Vol. 83, No. 5, pp. 741–770.
- [75] N. Morgan, H. Boulard, S. Greenberg and H. Hermansky (1995), "Stochastic perceptual models of speech", *IEEE Proc. Internat. Conf. Acoust. Speech Signal Process.*, pp. 397–400.
- [76] N. Morgan, S.-L. Wu and H. Boulard (1995), "Digit recognition with stochastic perceptual models", *Proc. Eurospeech'95, Madrid, Spain*, pp. 771–774.
- [77] L. Neumayer and M. Weintraub (1994), "Probabilistic optimum filtering for robust speech recognition", *IEEE Proc. Internat. Conf. Acoust. Speech Signal Process., Adelaide, Australia*, pp. 1-417–420.
- [78] M. Ostendorf, I. Bechwati and O. Kimball (1992), "Context modeling with the stochastic segment model", *IEEE Proc. Internat. Conf. Acoust. Speech Signal Process., San Francisco, CA*, pp. 389–392.
- [79] D.B. Paul, J.K. Baker and J.M. Baker (1991), "On the interaction between true source, training, and testing language models", *IEEE Proc. Internat. Conf. Acoust. Speech Signal Process., Toronto, Canada*, pp. 569–572.
- [80] M. Pavel and H. Hermansky (1994), "Temporal masking in automatic speech recognition", *J. Acoust. Soc. Amer.*, Vol. 95, No. 5, pp. 2876.
- [81] L.C.W. Pols, L.J.T. v.d. Kamp and R. Plomp (1969), "Perceptual and physical space of vowel sounds", *J. Acoust. Soc. Amer.*, Vol. 46, pp. 458–467.
- [82] L.C.W. Pols (1971), "Real-time recognition of spoken words", *IEEE Trans. Computers*, Vol. 20(C), pp. 972–978.
- [83] A.B. Poritz (1982), "Linear predictive hidden Markov models and the speech signal", *Proc. IEEE Internat. Conf. Acoust. Speech Signal Process., Paris, France*, pp. 1291–1294.
- [84] A.B. Poritz and A.G. Richter (1986), "On hidden Markov models in isolated word recognition", *Proc. IEEE Internat. Conf. Acoust. Speech Signal Process., Tokyo, Japan*, pp. 705–708.
- [85] J.E. Porter and S.F. Boll (1984), "Optimal estimators for spectral restoration of noisy speech", *Proc. IEEE Internat. Conf. Acoust. Speech Signal Process., San Diego, CA*, pp. 18.A.2.1.–18.A.2.4.
- [86] L.R. Rabiner (1989), "A tutorial on hidden Markov models and selected applications in speech recognition", *Proc. IEEE*, Vol. 77, No. 2, pp. 257–285.
- [87] S. Renals and M. Hochberg (1995), "Efficient search using posterior phone probability estimates", *Proc. IEEE Internat. Conf. Acoust. Speech Signal Process., Detroit, MI*, pp. 596–599.
- [88] M.D. Richard and R.P. Lippmann (1991), "Neural network classifiers estimate Bayesian a posteriori probabilities", *Neural Computation*, Vol. 3, pp. 461–483.
- [89] T. Robinson, L. Almeida, J.M. Boite, H. Boulard, F. Fallside, M. Hochberg, D. Kershaw, P. Kohn, Y. Konig, N. Morgan, J.P. Neto, S. Renals, M. Saerens and C. Wooters (1993), "A neural network based, speaker independent, large vocabulary, continuous speech recognition system: the Wernicke project", *Proc. Eurospeech'93, Berlin, Germany*, pp. 1941–1944.
- [90] T. Robinson and F. Fallside (1991), "A recurrent error propagation network speech recognition system", *Computer Speech and Language*, Vol. 5, pp. 259–274.
- [91] A.E. Rosenberg, C. Lee and F.K. Soong (1994), "Cepstral channel normalization techniques for HMM-based speaker verification", *Proc. Internat. Conf. on Spoken Language Processing, Yokohama, Japan*, pp. 1835–1838.
- [92] M.J. Russell, K.M. Ponting, S.M. Peeling, S.R. Browning, J.S. Bridle, R.K. Moore, I. Galiano and P. Howell (1990), "The ARM continuous speech recognition system", *Proc. IEEE Internat. Conf. Acoust. Speech Signal Process., Albuquerque, NM*, pp. 69–72.
- [93] R. Schwartz, S. Austin, F. Kubala, J. Makhoul, L. Nguyen, P. Placeway and G. Zavaliagos (1992), "New uses of the N-best sentence hypotheses within the BYBLOS speech recognition system", *Proc. IEEE Internat. Conf. Acoust. Speech Signal Process., San Francisco, CA*, pp. 1-1.4–1.7.
- [94] S. Seneff (1985), "A joint synchrony/mean-rate model of auditory speech processing", *J. Phonetics*, Vol. 16, No. 1, pp.55–76.
- [95] E. Singer and R. Lippmann (1992), "A speech recognizer using radial basis function neural networks in an HMM framework", *Proc. IEEE Internat. Conf. Acoust. Speech Signal Process., San Francisco, CA*, pp. 629–632.
- [96] J.M. Steeneken and D.A. Van Leeuwen (1995), "Multilingual assessment of speaker independent large vocabulary speech-recognition systems: The SQALE project (speech recognition quality assessment for language engineering)", *Proc. Eurospeech'95, Madrid, Spain*, pp. 1271–1274.
- [97] S.S. Stevens (1957), "On the psychophysical law", *Psychol. Rev.*, Vol. 64, No. 1, pp. 153–181.
- [98] J.C. Stevens and J.W. Hall (1966), "Brightness and loud-

- ness as functions of stimulus duration”, *Perception and Psychophysics*, pp. 319–327.
- [99] T. Stockham, T. Cannon and R. Ingerbretsen (1975), “Blind deconvolution through digital signal processing”, *Proc. IEEE*, Vol. 63, pp. 678–692.
- [100] C.C. Tappert and S.K. Das (1978), “Memory and time improvements in a dynamic programming algorithm for matching speech patterns”, *IEEE Trans. Acoust. Speech Signal Process.*, Vol. 26, pp. 583–586.
- [101] J. Tebelskis, A. Waibel, B. Petek and O. Schmidbauer (1991), “Continuous speech recognition using linked predictive neural networks”, *Proc. IEEE Internat. Conf. Acoust. Speech Signal Process., Toronto, Canada*, pp. 61–64.
- [102] A. Waibel, T. Hanazawa, G. Hinton, K. Shikano and K. Lang (1988), “Phoneme recognition using time-delay neural networks”, *Proc. IEEE Internat. Conf. Acoust. Speech Signal Process., New York, NY*, pp. 107–110.
- [103] C.J. Wellekens (1987), “Explicit time correlation in hidden Markov models for speech recognition”, *Proc. IEEE Internat. Conf. Acoust. Speech Signal Process., Dallas, TX*, pp. 384–386.
- [104] E. Zwicker (1975), “Scaling”, in *Handbook of Sensory Physiology*, ed. by Keidel and Neff (Springer, Berlin, Vol. 3), pp. 401–448.
- [105] V. Zue (1985), “The use of speech knowledge in automatic speech recognition”, *Proc. IEEE*, Vol. 73, No. 11, pp. 1602–1615.
- [106] H. Boulard, H. Hermansky and N. Morgan (1996), “Copernicus and the ASR challenge – Waiting for Kepler”, *Proc. ARPA Speech Recognition Workshop, Arden House, NY, 18–21 February 1996*.