TOWARDS ROBUST AND ADAPTIVE SPEECH RECOGNITION MODELS

HERVE BOURLARD*, SAMY BENGIO†, AND KATRIN WEBER‡

Abstract.

In this paper, we discuss a family of new Automatic Speech Recognition (ASR) approaches, which somewhat deviate from the usual ASR approaches but which have recently been shown to be more robust to nonstationary noise, without requiring specific adaptation or "multi-style" training. More specifically, we will motivate and briefly describe new approaches based on multi-stream and subband ASR. These approaches extend the standard hidden Markov model (HMM) based approach by assuming that the different (frequency) streams representing the speech signal are processed by different (independent) "experts", each expert focusing on a different characteristic of the signal, and that the different stream likelihoods (or posteriors) are combined at some (temporal) stage to yield a global recognition output. As a further extension to multi-stream ASR, we will finally introduce a new approach, referred to as HMM2, where the HMM emission probabilities are estimated via state specific feature based HMMs responsible for merging the stream information and modeling their possible correlation.

Key words. Robust speech recognition, hidden Markov models, subband processing, multi-stream processing.

1. Introduction. Current automatic speech recognition systems are based on (context-dependent or context-independent) phone models described in terms of a sequence of hidden Markov model (HMM) states, where each HMM state is assumed to be characterized by a stationary probability density function. Furthermore, time correlation, and consequently the dynamic of the signal, inside each HMM state is also usually disregarded (although the use of delta and delta-delta features can capture some of this correlation). Consequently, apart from the dependencies captured via the topology of the HMM model, most time dependencies are usually very poorly modeled.[1] Ideally, we want to design a particular HMM able to accommodate multiple time-scale characteristics so that we can capture phonetic properties, as well as syllable structures, which seem to have many attractive properties [9], including invariants that are more robust to noise.

[*] Dalle Molle Institute for Perceptual Artificial Intelligence (IDIAP), and Swiss Federal Institute of Technology at Lausanne (EPFL), 4, Rue du Simplon, CH-1920 Martigny, Switzerland, bourlard@idiap.ch.

[†] Dalle Molle Institute for Perceptual Artificial Intelligence (IDIAP), 4, Rue du Simplon, CH-1920 Martigny, Switzerland, bengio@idiap.ch.

[‡] Dalle Molle Institute for Perceptual Artificial Intelligence (IDIAP), and Swiss Federal Institute of Technology at Lausanne (EPFL), 4, Rue du Simplon, CH-1920 Martigny, Switzerland, weber@idiap.ch.

[1] This problem is not specific to the fact that phone models are generally used. Whole word models, or syllable models, built up as sequences of HMM states will suffer from exactly the same drawbacks, the only potential advantage of moving towards "larger" units being that one can then have more word (or syllable) specific distributions (usually resulting in more parameters and an increased risk of undersampled training data). Consequently, building an ASR system simply based on syllabic HMMs will not alleviate the limitations of the current recognizers since those models will still be based on the short-term piecewise stationary assumptions mentioned above.

For example, acoustic features such as the modulation spectrogram[2] exhibit some correlation with syllabic features and can be used to improve state-of-the-art ASR systems [33]. It is, however, clear that those different time-scale features will also exhibit different levels of stationarity and will require different HMM topologies to capture their dynamics.

There are many potential advantages to such a multi-stream approach, including:

1. The definition of a principled way to merge different temporal knowledge sources such as acoustic and visual inputs, even if the temporal sequences are not synchronous and do not have the same data rate – see [28] and [29] for further discussion about this.
2. The possibility to incorporate multiple time resolutions (as part of a structure with multiple unit lengths, such as phone and syllable).
3. Multiband-based ASR [6, 14] involving the independent processing and combination of partial frequency bands is a very particular case of multi-stream recognition. Although this will not be explicitly discussed here, there are many potential advantages to this multiband approach, including (i) better robustness to speech impaired by narrowband noise, and (ii) possibility to apply different time/frequency tradeoffs and different recognition strategies in the subbands.

In the following, we will not discuss the underlying algorithms ("complex" variants of Viterbi decoding, if one wants to take the possible asynchrony into account), nor detailed experimental results (see [11] for recent results). Instead, we will mainly discuss different combination strategies pointing towards the same formalism.

2. Psycho-Acoustic Evidence.

It seems to me that what can happen in the future is... that experiments get harder and harder to make, more and more expensive... and scientific discovery gets slower and slower. (Richard Feynman, 1918-1988, *The Character of Physical Law*, Cambridge, MA, p.172.)

2.1. Product of errors rule and its interpretation.

The work of Fletcher and his colleagues (see the insightful review of his work in [1]) suggests that human decoding of the linguistic message is based on decisions within narrow frequency subbands that are processed quite independently of each other. Empirical evidence suggests that the combination of decisions from these subbands is done at some intermediate level and in such a way that the global error rate is equal to the product of error rates in the subbands. In other words, if we have two frequency bands (streams) c_1 and c_2, and each of them is respectively yielding a probability of error (error rate) $e(q_j|x^1)$ and $e(q_j|x^2)$ for a particular class q_j and an input pattern $x = \{x^1, x^2\}$, where x^1 and x^2 represent the output features of the two frequency streams[3], the total error rate $e(q_j|x^1, x^2)$ resulting from the

[2] Initially proposed as a way to assess room acoustics [16].

[3] Since we decided not to deal with the temporal constraints, this notation is over-simplified. In the case of temporal sequences, x^1 and x^2 will be sequences (possibly of different lengths and different

simultaneous use of the two streams is given by:

$$e(q_j|x^1, x^2) = e(q_j|x^1)e(q_j|x^2) \ . \tag{2.1}$$

Although this conclusion is often questioned by the scientific community[4], it is probably not worth arguing too long about it since it is pretty clear that (2.1) is anyway the optimal rule to obtain the best performance out of a (possibly noisy) multi-stream system (but requiring the *perfect* knowledge of the noisy stream). Moreover, a similar rule can usually explain some of the empirical observations in audio-visual processing (see, e.g., [24] and [20]).

Although pretty simple, rule (2.1) is not always easy to interpret (and even less for engineers!). So let us have a closer look at it. Since the probability of being correct whenever we assign a particular observation x to a class q is equal to the *a posteriori probability* $P(q|x)$ (i.e., the probability of error is equal to $1 - P(q|x)$, see [8], page 12)[5], rule (2.1) can also be written as:

$$\begin{aligned} e(q_j|x^1, x^2) &= (1 - P(q_j|x^1))(1 - P(q_j|x^2)) \\ &= 1 - \sum_{k=1}^{2} P(q_j|x^k) + \prod_{k=1}^{2} P(q_j|x^k) \end{aligned} \tag{2.2}$$

where $P(q_j|x^k)$ denotes the class posterior probabilities obtained for the k-th input stream. Rewriting (2.2) in terms of (total) correct recognition probability ($P(q_j|x^1, x^2) = 1 - e(q_j|x^1, x^2)$), we have:

$$P(q_j|x^1, x^2) = \sum_{k=1}^{2} P(q_j|x^k) - \prod_{k=1}^{2} P(q_j|x^k) \tag{2.3}$$

In the case of K streams, the above expression will have $2^K - 1$ terms, containing all possible stream combinations.

These expressions are quite reasonable since they also reflect a standard property of probabilities of joint events.[6] Actually, this *product of errors rule* tells us that the probability of correct classification on human full-band hearing is equal to the probability that there is correct (human) classification in *any* subband. Consequently, this also means that human hearing seems capable of processing numerous bands and selecting the one that gives correct recognition.

The resulting (very simple but nonlinear) product of errors function is illustrated in Figure 1 for all possible values of $P(q_j|x^1)$ (horizontal axis) and

rates) of features, and q_j will be an HMM.

[4] Since the relevant Fletcher experiments were done (i) with nonsense syllables only, and (ii) using high-pass or low-pass filters (i.e., two streams) only, it is not clear whether or not this is an accurate statement for disparate bands in continuous speech.

[5] See Section 3 for further evidence.

[6] The probability of union of two events $P(A \text{ or } B) = P(A) + P(B) - P(A, B)$, which is also equal to $P(A) + P(B) - P(A)P(B)$ if events A and B are independent. Indeed, in estimating the proportions of a sequence of trials in which A and B occur, respectively, one counts twice those trials in which both occur.

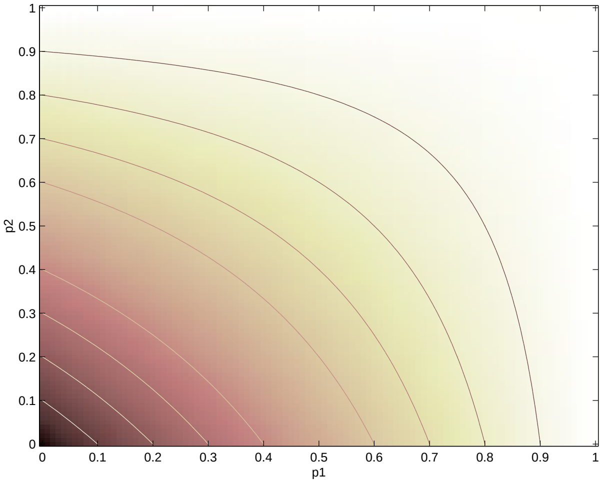

FIG. 1. *"Optimal" classification strategy based on two (independent) observation streams yielding posterior probabilities $P(q_j|x^1)$ and $P(q_j|x^2)$. The grey level represents the "total" probability of correct recognition (with white corresponding to the maximum probability), and the different curves represent the equal recognition probability curves (as a function of $P(q|x^1)$ and $P(q|x^2)$) above which the probability of correct recognition will be higher than a prescribed value.*

$P(q_j|x^2)$ (vertical axis). From this figure, it is interesting to note how much flexibility an "optimal" multi-stream system potentially has in keeping the (total) probability of correct recognition above a certain threshold, even if one of the streams is extremely noisy (and yields high error rates). This can indeed be measured by the area above a given equal recognition rate curve. For example, for $P(q_j|x^1, x^2) = 0.9$, nearly one third of the space is available! It is clear that this property cannot be achieved by using the usual product of likelihoods, where if one of the likelihoods is poorly estimated, the whole product is deteriorated.[7]

This conclusion remains valid for more than two streams. Actually, it can even be shown that the area above a given equal error rate (multi-dimensional) surface is growing exponentially with the number of streams. To make the link easier with what will come in the sequel of this paper, it is easy to show that, in the case of three input streams, (2.3) becomes:

$$\begin{aligned} P(q_j|x^1, x^2, x^3) = &\sum_{k=1}^{3} P(q_j|x^k) + \prod_{k=1}^{3} P(q_j|x^k) \\ &- \sum_{\ell>k=1}^{3} P(q_j|x^k)P(q_j|x^\ell) \end{aligned} \tag{2.4}$$

[7] On top of the fact that it is usually difficult to compare/combine likelihoods computed from features in different spaces, possibly of different dimensions (since likelihoods, as usually computed (assuming Gaussian densities with diagonal covariance matrices), are "dimensional", i.e., depend on the dimension of the feature space).

Obviously, this reflects a “perfect” world. In actual engineering systems though the posterior probabilities $P(q_j|x^k)$ will have to be estimated on the basis of a set of parameters Θ, and, in the case of two streams, (2.3) should be written:

$$\begin{aligned} & P(q_j|x^1, x^2, \Theta) \\ & = \sum_{k=1}^{2} P(q_j|x^k, \Theta) - \prod_{k=1}^{2} P(q_j|x^k, \Theta) \end{aligned} \tag{2.5}$$

Figure 1 does not change, but the position in the space depends on Θ, as well as on the different stream features. Ideally, robust training and adaptation should be performed in the Θ space to guarantee that $P(q_j|x^1, x^2, \Theta)$ is always above a certain threshold, or to directly maximize (2.5). In the following, we discuss approaches going in that direction.

2.2. Discussion. The above analysis allows us to draw a few conclusions and to design the features of an “optimal” ASR system:

1. Human hearing performs combination of frequency streams according to the product of errors rule discussed above. In this case (and assuming that the subbands are independent, which is false), correct classification of any subband is empirically equivalent to correct full-band classification. *In subband-based ASR systems, this means that we should design the system and the training criterion to maximize the classification performance on subbands, while also making sure that the subband error rates are independent.*
2. As a direct consequence of the above, it is also obvious that the more subbands we use, the higher the full-band correct classification rate will be. As done in human hearing, *ASR systems should thus use a large number of subbands* to have a better chance to increase recognition rates. It is interesting to note here that this trend has recently been followed in [15].
3. In order to estimate the reliability of each stream, *ASR systems should be able to estimate subband posteriors as accurately as possible.* We will show in the next section that this is not impossible.
4. If ASR systems can reliably estimate local posteriors, we can implement the product of errors rule, which should guarantee the minimum of errors (if the above conditions are satisfied). Furthermore, each time we improve the classification rate in *any* subband, the recognition rate should improve.

3. Estimating Posteriors. *The purpose of models is not to fit the data but to sharpen the questions.* (Samuel Karlin, 1923-, 11th R.A. Fisher Memorial Lecture, Royal Society, 20 April 1983.)

From the discussion above, it seems clear that we should work on the basis of a posteriori probabilities[8]. Given that we often work in the framework of hybrid

[8] Which are known, anyway, to yield the minimum error rate solution.

HMM/ANN systems [5] (using artificial neural networks (ANN) for estimating local posterior probabilities which are transformed into scaled likelihoods used as HMM emission probabilities), and although some of the arguments below will also be valid for likelihood-based systems, we will focus our discussion on posteriors.

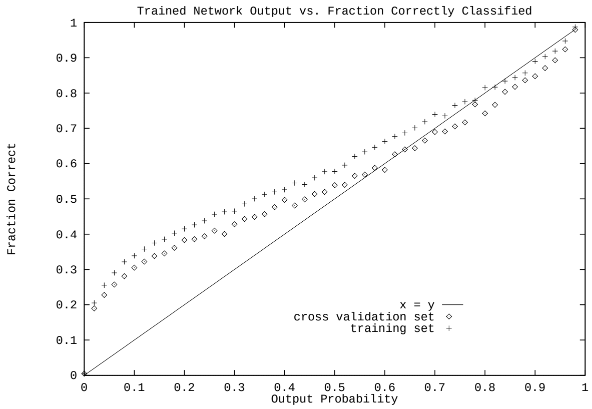

FIG. 2. *It is possible to generate "good" posterior probabilities out of a neural network, and these are indeed good measures of the probability of being correct. This plot was generated on real speech data by collecting statistics over the acoustic parameters from 1750 Resource Management speaker-independent training sentences and 500 cross-validation sentences (not used for training, but for which correct classification was known).*

As initially reported in [5], Figure 2 illustrates the fact that ANN can reliably estimate local posterior probabilities $P(q_j|x)$. Indeed, recalling the properties of posterior probabilities discussed in the previous section, good estimates of posterior probabilities should also be a measure of the fraction of correct classification. Consequently, when representing the correct classification rate as a function of the posterior probabilities as estimated at the output of a neural network, the ideal Bayes (posterior-based) classifier would yield a diagonal, which is quite the case for both the training data and the cross-validation data (not used for training, but for which correct classification was known).

Dividing these local posterior probabilities by the prior probabilities $P(q_j)$ as estimated on the training set, yields scaled local likelihoods that can be used to compute [12]

$$\frac{P(M|X)}{P(M)} = \frac{P(X|M)}{P(X)} \tag{3.1}$$

where M represents a complete HMM (modeling a particular sub-unit, a word, or a sentence) composed of several units computing $\frac{p(q_j|x)}{p(q_j)}$, and X an observation sequence associated with M. This can then be simply multiplied (as in usual HMMs) by $P(M)$ to include external knowledge sources (such as a language model).

4. Multi-Stream and Mixture of Experts. From an engineering perspective, one way to introduce the multi-stream formalism in a pattern classification (such as ASR) task is to use the approach of mixture of experts, as proposed in the framework of neural networks [4]. The general idea of mixture of experts is to process the (same) input space according to different linear or nonlinear (neural network) functions ("experts"), and to combine the outputs of each expert according to a weighted sum, and where the weights also result from a (linear or nonlinear) function of the input pattern x.

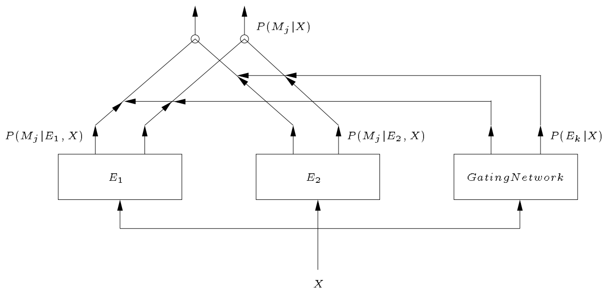

FIG. 3. *Posterior-based mixture of experts. Experts (e.g., neural networks) are extracting their own posterior estimates, which are then combined through weights also estimated (by the "gating network") from the data. These weights could also be adapted online.*

Typically, this approach (as for HMMs) can be formulated in terms of latent variables (where the missing variable is the expert sequence). As illustrated in Figure 3, let M represent the hypothesized model (HMM) associated with an input sequence X. If $\mathcal{E} = \{E_1, \ldots, E_k, \ldots, E_K\}$ represents a set of *mutually exclusive and exhaustive* experts[9] (and where $P(E_k)$ is defined as the probability that E_k is the most reliable expert), then $P(M|X)$ can be estimated as:

$$\begin{aligned} P(M|X) &= \sum_{k=1}^{K} P(M, E_k|X) \\ &= \sum_{k=1}^{K} P(M|E_k, X)P(E_k|X) \\ &\simeq \sum_{k=1}^{K} P(M^k|X^k)P(E_k|X) \qquad (4.1) \end{aligned}$$

where X^k represents the respective inputs of expert/function E_k [10], M^k the model for the speech unit M used to process X^k, and $P(E_k|X)$ the (relative) reliability

[9] As discussed later, the initial multi-stream approach (Section 5.1) was not using strictly exhaustive experts since they did not cover all possible stream combinations. The full combination approach, as discussed in Section 5.3, will actually use all possible combinations.

[10] In the case of multi-stream inputs, X^k will typically be a subset of X (containing the features relative to E_k).

of expert E_k given the whole input.[11] The approximations in (4.1) result from the assumptions that (i) the probability of a model M given a particular expert E_k is only estimated from the sub-model M^k associated with the expert, and (ii) that expert-specific model is only looking at its specific input features. The segment-based posteriors in (4.1) can be computed as briefly recalled in Section 3.

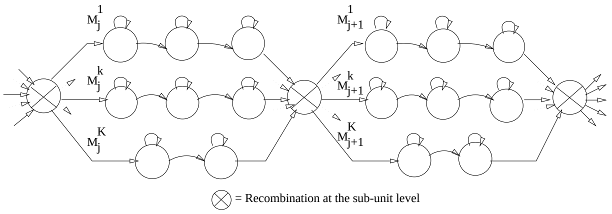

FIG. 4. *General form of a K-streams recognizer with anchor points between speech units (to force synchrony between the different streams). Note that the model topology is not necessarily the same for the different sub-systems.*

Ideally, as discussed in [6, 14, 29] and illustrated in Figure 4, the expert combination presented above should take place at the level of M, i.e., at the level of the particular (non-emitting) states denoted “$\otimes$”. However, this is not trivial and will often require a significant adaptation of the recognizer. It is only in the case of segment likelihoods combination (by products) that one can develop a tractable solution to this optimization problem. Indeed, in this particular case, it is easy to show that the product of segment-based, expert-specific, likelihoods can be distributed through local likelihood products of an equivalent 1st order HMM, possibly after some modification of the transition probabilities [32]. This algorithm, referred to as “HMM combination”, is an adaptation of the HMM decomposition algorithm presented in [30].

In the case of more complex (non linear) combination criteria, like in the case of mixture of experts or the approach discussed is section 5 (related to the mixture of experts model and the psycho-acoustic evidence discussed in Section 2), HMM combination/decomposition is no longer a tractable solution. Other approaches based on the 2-level dynamic programming algorithm or using (4.1) to rescore an N-best list of hypotheses (providing us with a set of possible segmentation/anchor points) have then to be used.

Although it is clear that:

1. The empirical results discussed in Section 2 were obtained on the basis of segments (non-sense syllables),
2. only the segment level combination can allow for asynchrony between the streams[12],

[11] Since, as illustrated in Figure 4, each sequence X^k will be processed with a different/specific HMM.

[12] Although not using the nonlinear (optimal?) combination functions discussed in this paper, preliminary results presented in [6, 14] suggested that asynchrony was not a major factor — see, though, [21] and [29] for further discussion about this.

we will mainly focus, in the sequel of this paper, on the combination at the state level.

5. Multiband-based ASR with Latent Variables.

5.1. General Formalism. As a particular case of multi-stream processing, we have been investigating an ASR approach based on independent processing and combination of frequency subbands. The general idea, as illustrated in Fig. 5, is to split the whole frequency band (represented in terms of critical bands) into a few subbands on which different recognizers are independently applied. The resulting probabilities are then combined for recognition later in the process at some segmental level (here we consider the state level). Starting from critical bands, acoustic processing is now performed independently for each frequency band, yielding K input streams, each being associated with a particular frequency band.

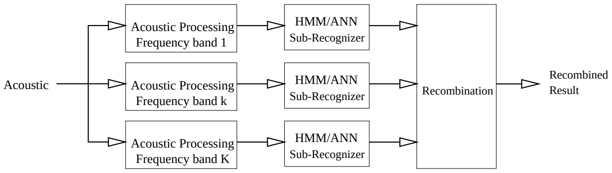
Acoustic
Acoustic Processing Frequency band 1
Acoustic Processing Frequency band k
Acoustic Processing Frequency band K
HMM/ANN Sub-Recognizer
HMM/ANN Sub-Recognizer
HMM/ANN Sub-Recognizer
Recombination
Recombined Result

FIG. 5. *Typical multiband-based ASR architecture. In multiband speech recognition, the frequency range is split into several bands, and information in the bands is used for phonetic probability estimation by independent modules. These probabilities are then combined for recognition later in the process at some segmental level.*

In this case, each of the K sub-recognizers (streams) is now using the information contained in a specific frequency band $X^k = \{x_1^k, x_2^k, \ldots, x_n^k, \ldots, x_N^k\}$, where each x_n^k represents the acoustic (spectral) vector at time n in the k-th stream. In (4.1), $P(M^k|X^k)$ represents the a posteriori probability of a sub-model M^k (k-th frequency band model for M) and can be estimated from local posteriors $P(q_j^k|x_n^k)$ (e.g., estimated at the output of an ANN), where q_j^k denotes a state j of model M^k). $P(E_k|X)$ represents the "reliability" of expert E_k, working on the k-th frequency band, and can be estimated in different ways (e.g., based on SNR).

As discussed in the previous section, combination at the segment level according to the criteria discussed here is not easy. However, combination at the HMM-state level, by combining local posteriors $P(q_j^k|x_n^k)$, can be done in many ways [6], including untrained linear or trained linear (e.g., as a function of automatically estimated local SNR) functions, as well as trained nonlinear function (e.g., by using a neural network). This is pretty simple to implement and amounts to performing a standard Viterbi decoding in which local (log) probabilities are obtained from a linear or nonlinear combination of the local subband probabilities. For example, in the initial subband-based ASR, local posteriors $P(q_j|x_n)$

(or scaled likelihoods) were estimated according to:

$$P(q_j|x_n) = \sum_{k=1}^{K} w_k P(q_j|x_n^k, \Theta_k) \tag{5.1}$$

where, in our case, each $P(q_j|x_n^k, \Theta_k)$ is computed with a band-specific ANN of parameters Θ_k and with x_n^k (possibly with temporal context) at its input. The weighting factors can be assigned a uniform distribution (already performing very well [6]) or be proportional to the estimated SNR. Over the last few years, several results were reported showing that such a simple approach was usually quite robust to band limited noise.

In Section 5.3 below, we discuss a new approach that was recently developed at IDIAP, and presented in [3, 11, 23], and show (i) how it significantly enhances the baseline multiband approach, and (ii) how it relates to the above discussions (and psycho-acoustic evidence).

5.2. Motivations and Drawbacks. The multiband approach has several potential advantages, which are briefly discussed here.

Better robustness to band-limited noise — The signal may be impaired (e.g., by noise, stream characteristics, reverberation,...) only in some specific frequency bands. When recognition is based on several independent decisions from different frequency subbands, the decoding of a linguistic message need not be severely impaired, as long as the remaining clean subbands supply sufficiently reliable information. This was confirmed by several experiments (see, e.g., [6]). Surprisingly, even when the combination is simply performed at the HMM state level, it is observed that the multiband approach is yielding better performance and noise robustness than a regular full-band system.[13]

Similar conclusions were also observed in the framework of the missing feature theory [19, 22]. In this case, it was shown that, *if one knows the position of the noisy features*, significantly better classification performance could be achieved by disregarding the noisy data (using marginal distributions) or by integrating over all possible values of the missing data conditionally on the clean features — See Section 5.3 for further discussion about this. In the multiband approach, we do not try to explicitly identify the noisy band (and to disregard it). Instead, we process all the subbands independently (to avoid “spreading” the noise across all components of the feature vector or in the local probability estimate) and recombine them according to a particular weighting scheme that should de-emphasize (or cancel out) the noisy bands.

[13] It could however be argued that, in this case, the multiband approach boils down to a regular full-band recognizer in which several likelihoods of (assumed) independent features are estimated and multiplied together to yield local likelihoods (since, in likelihood based systems, expected values for the full-band is the same than the concatenated expected values of subbands). This is however not true when using posterior based systems (such as hybrid HMM/ANN systems) where the subbands are presented to different nets that are independently trained in a discriminant way on each individual subband. Finally, as discussed in this paper, we also believe that the combination criterion should be different than a simple product of (scaled) likelihoods or posteriors.

Better modeling — As for a regular full-band system, it was shown in [6] that all-pole modeling was significantly improving the performance of multiband systems. However, as an additional advantage of the subband approach, it can be shown or argued that:

1. This all-pole modeling may be more robust if performed on several subbands (low dimensional spaces) than on the full-band signal [27].
2. Since the dimension of each (subband) feature space is smaller, it is easier to estimate reliable statistics (resulting in a more robust parameterization).

Stream asynchrony — Transitions between more stationary segments of speech do not necessarily occur at the same time for the different frequency bands [21], which makes the piecewise stationary assumption more fragile. The subband approach may have the potential of relaxing the synchrony constraint inherent in current HMM systems.

Stream specific processing and modeling — Different recognition strategies might ultimately be applied in different subbands. For example, different time/frequency resolution tradeoffs could be chosen (e.g., time resolution and width of analysis window depending on the frequency subband). Finally, some subbands may be inherently better for certain classes of speech sounds than others.

Major objections and drawbacks — There are a few, related, drawbacks to this multiband approach [21]:

1. One of the common objections to this separate modeling of each frequency band has been that important information in the form of correlation between bands may be lost. Although this may be true, several studies [21], as well as the good recognition rates achieved on small frequency bands [10, 15], tend to show that most of the phonetic information is contained in each frequency band (possibly provided that we have enough temporal information)[14].
2. To define and independently process frequency bands, it is obviously necessary to start from spectral coefficients (critical bands), which, however, are not orthogonal and do not permit competitive performance for clean speech. In standard ASR systems, these coefficient are typically orthogonalized using a DCT (cepstral) transformation. Even in the case of ANN probability estimation (where ANN is supposed to extract and model the correlation across coefficients), it has been observed that orthogonalization of the features still helped a bit. However, in the case of narrowband additive noise, we obviously want to subtract as much as possible of the noise before the DCT transform to avoid spreading the noise across all the feature components. For subband ASR systems, a

[14] And, indeed, the discussion in Section 2, as well as many other psycho-acoustic experiments, seem to suggest that human hearing can actually extract a lot of phonetic/syllabic information from band limited signals.

partial but effective solution to this problem consists in performing an independent DCT in each subband [6, 26].

Alternative solutions to this problem have recently been proposed in which it is attempted to decorrelate as much as possible the filter-bank energies — see, e.g., [18, 7, 25]. This is usually obtained by performing some kind of temporal filtering (and, consequently, spreading the possible noise over time instead of over frequency) or frequency filtering (and consequently spreading the possible noise over a limited frequency range only).

3. As opposed to the empirical evidence discussed in Section 2, the initial subband-based ASR system did not make use of all possible subband combinations. This will be fixed by the method presented next.

5.3. Full Combination Subband ASR. Following the above developments and discussions, it seems reasonable to assume that a subband ASR system should simultaneously deal with all the $L = 2^K$ possible subband combinations S^ℓ (with $\ell = 1, \ldots, L$, including the empty set[15]) resulting from an initial set of K frequency (critical) bands x^k. However, while it is pretty easy to quickly estimate any subband likelihood or marginal distribution when working with Gaussian or multi-Gaussian densities [19], this is harder when using ANN to estimate posterior probabilities. In this latter case, indeed, it would be necessary to train (and run, during recognition) 2^K neural networks, which would become very quickly intractable.

In the following, we briefly present the solution recently proposed in [11] and [23], and discuss its relationships with the themes developed in the current paper.

Ideally, we would thus like to compute the posterior probabilities for each of the $L = 2^K$ possible combinations S_n^ℓ (including all possible single bands, pairs of bands, triples, etc) of the K subbands x_n^k. Indeed, since we do not know a priori where the noise is located, we should integrate over all possible positions[16]. Using the formalism of mixture of experts, we can thus write:

$$\begin{aligned} P(q_j|x_n,\Theta) &= \sum_{\ell=1}^{L} P(q_j, E_\ell|x_n,\Theta) \\ &= \sum_{\ell=1}^{L} P(q_j|E_\ell, x_n,\Theta)P(E_\ell|x_n) \\ &= \sum_{\ell=1}^{L} P(q_j|S_n^\ell,\Theta_\ell)P(E_\ell|x_n) \qquad (5.2) \end{aligned}$$

[15] Which would correspond to the case where all the bands are unrealiable. In this case, the best posterior estimate is the prior probability $P(q_j)$, and one of the L terms in the following equations will contain only this prior information.

[16] This amounts to assuming that the position of the noise or, in other words, the position of the reliable frequency bands, is a hidden (latent) variable on which we will integrate to maximize the posterior probabilities (in the spirit of the EM algorithm).

where Θ represents the whole parameter space, while Θ_ℓ denotes the set of (ANN) parameters used to compute the subband posteriors. Of course, implementation of (5.2) requires the training of L neural networks to estimate all the posteriors $P(q_j|S_n^\ell, \Theta_\ell)$ that have to be combined according to a weighted sum, with each weight representing the relative reliability of a specific set of subbands. In the case of stationary interference, this reliability could be estimated on the basis of the average (local) SNR in the considered set. Alternatively, it could also be estimated as the probability that the local SNR is above a certain threshold, and where the threshold has been estimated to guarantee a prescribed recognition rate (e.g., lying above a certain equal recognition rate curve in Figure 1) [3].

Typically, training of the L neural nets would be done once and for all on clean data, and the recognizer would then be *adapted* online simply by adjusting the weights $P(E_\ell|x_n)$ (still representing a limited set of L weights) to increase the global posteriors. This adaptation could be performed by online estimation of the SNR or by an online version of the EM (deleted-interpolation) algorithm. Although this approach is not really tractable, it has the advantage of avoiding the independence assumption between the subbands of a same set, as well as allowing any DCT transformation of the combination before further processing. Consequently, this combination, referred to as *Full Combination,* was actually implemented [10] for the case of four frequency subbands (each containing several critical bands), thus requiring the training of 16 neural nets, and used as an "optimal" reference point.

An interesting approximation to this "optimal" solution though consists in simply train one neural network per subband for a total of K models, and to approximate all the other subband combination probabilities directly from these. In other words, re-introducing the independence assumption[17] between subbands, subband combination posteriors would be estimated as [10, 11]:

$$P(q_j|S_n^\ell, \Theta_\ell) = P(q_j) \prod_{k \in S^\ell} \frac{P(q_j|x_n^k, \Theta_k)}{P(q_j)} \tag{5.3}$$

Experimental results obtained from this approximated Full Combination approach in different noisy conditions are reported in [10, 11], where the performance of this above approximation was also compared to the "optimal" estimators (5.2). Interestingly, it was shown that this independence assumption did not hurt us much and that the resulting recognition performance[18] was similar to the performance obtained by training and recombining all possible L nets (and significantly better than the original subband approach). In both cases, the recognition rate and the robustness to noise were greatly improved compared to the initial subband approach (5.1). This further confirms that we do not seem to lose "critically" important information when neglecting the correlation between bands.

[17] Actually, it is shown in [10, 11] that we only need to introduce a weak (conditional) independence assumption.

[18] Obtained on the Numbers'95 database, containing telephone-based speaker independent free format numbers, on which NOISEX noise was added.

Finally, it is particularly interesting to note here that using (5.3) in (5.2) yields something very similar to the "optimal" product of errors rule (2.4) observed empirically:

$$P(q_j|x_n,\Theta) = \sum_{\ell=1}^{L} \frac{P(E_\ell|x_n)}{C_\ell} \prod_{k\in S^\ell} P(q_j|x_n^k,\Theta_k) \tag{5.4}$$

with $C_\ell = P^{(n_\ell-1)}(q_j)$, and n_ℓ being the number of subbands in S^ℓ. In [10], it is shown that this normalization factor is important to achieve good performance. This Full Combination rule thus takes exactly the same form than the product of errors rule [such as (2.4) or (2.5)], apart from the fact that the weighting factors are different. In (5.4), the weigthing factors can be interpreted as (scaled) probabilities estimating the relative reliability of each combination, while in the product of errors rule these are simply equal to $+1$ or -1. Another difference is that the product of errors rule involves $2^K - 1$ terms while the Full Combination rule involves 2^K terms, one of them representing the contribution of the prior probability.

In the next section, we discuss a further extension of this approach where the segmentation into subbands is no longer done explicitly, but is achieved dynamically over time, and where the integration over all possible frequency segmentations is part of the same formalism.

6. HMM2: Mixture of HMMs. All HMM emission probabilities discussed in the previous models are typically modeled through Gaussian mixtures or artificial neural networks. Also, in the multiband based recognizers discussed above, we have to decide *a priori* the number and position of the subbands being considered. As also briefly discussed above, it is not always clear what the "optimal" recombination criterion should be. In the following, we introduce a new approach, referred to as HMM2, where the emission probabilities of the HMM (now referred to as "temporal HMMs") are estimated through a secondary, state-dependent, HMM (referred to as "feature HMMs") specifically working along the feature vector. As briefly discussed below (see references such as [2] and [31] for further detail), this model will then allow for dynamic (time and state dependent) subband (frequency) segmentation as well as "optimal" recombination according to a standard maximum likelihood criterion (although other criteria used in standard HMMs could also be used).

In HMM2, as illustrated in Figure 6, each temporal feature vector x_n is considered as a fixed length sequence of S components $x_n = (x_n^1, .., x_n^S)$, which is supposed to have been generated at time n by a specific feature HMM associated with a specific state q_j of the temporal HMM. Each feature HMM state r_l is thus emitting individual feature components x_n^s, whose distributions are modeled by, e.g., one dimensional Gaussian mixtures. The feature HMM thus looks at all possible subband segmentations and automatically performs the combination of the likelihoods to yield a single emission probability. The resulting emission probability can then be used as emission probability of the temporal HMM. As an alternative, we can also use the resulting feature segmentation in multiband systems or as additional acoustic features features in a standard ASR system. Indeed, if HMM2 is applied to the spectral domain, it is expected that the feature HMM will "segment" the feature vector into piecewise stationary spectral regions, which could thus follow spectral peaks (formants) and/or spectral valley regions.

In the example illustrated in Figure 6, the HMM2 is composed of a temporal HMM that handles sequences of features through time, and feature HMMs assigned to the different temporal HMM states. The temporal HMM is composed of 3 left-to-right connected states (q_1, q_2 and q_3), while the state-specific feature HMM is composed of 4 ("top-down") states (r_1, r_2 r_3 and r_4). Although not reflected in Figure 6, each feature HMM $\{r_1, r_2, r_3, r_4\}$ is specific to a temporal HMM state (emission probability distribution), with different parameters, and possibly different HMM topologies. More formally, as done in [2] and [17], the feature state should have been denoted r_j^k, with k representing the associated temporal state index and j the feature state index.

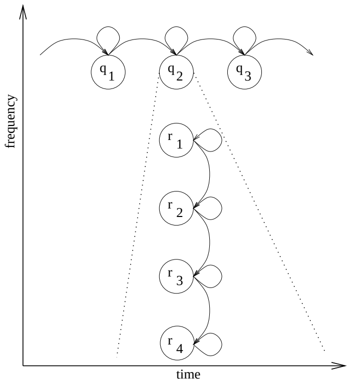

FIG. 6. *HMM2: the emission distributions of the temporal HMM are estimated by secondary, state-specific, feature HMMs.*

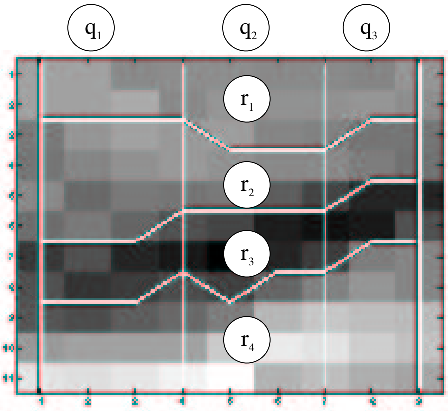

FIG. 7. *Frequency filtered filterbanks and HMM2 resulting (Viterbi) segmentation for a test example of phoneme* "w".

Of course, the topology of the feature HMM, extracting the correlation information withing feature vectors, could take many forms, including ergodic HMMs and/or topologies with a number of states larger the number of feature components, in which case "high-order" correlation information could be modeled. In the following though, we constrained the feature HMM to a strictly "top-down" topology. Moreover, since we were interested in extracting information in the spectral domain and in possible relationships with multiband ASR systems, we considered features in the spectral domain. Each of the feature HMM states is then supposed to model one of the K frequency bands, where the positions and bandwidths of these bands are determined dynamically.

In [2], we introduced an EM algorithm to jointly train all the parameters of such an HMM2 in order to maximize the data likelihood. This derivation is based on the fact that an HMM is a special kind of mixture of distributions, and therefore HMM2, as a mixture of HMMs, can be considered as a more general kind of mixture distribution. During decoding, the Viterbi algorithm is used to find the path through the HMM2 which best explains the input data. Local state likelihoods of the temporal HMM can however be estimated using either Viterbi or the complete likelihood calculation, summing over all possible paths through the feature HMM:

$$p(x_n|q_j) = \sum_R P(r_0|q_j) \prod_{s=1}^{S} p(x_n^s|r_l, q_j) P(r_l|r_{l-1}, q_j) \tag{6.1}$$

where q_j is the temporal HMM state at time n, r_l the feature HMM state at feature s, R the set of all possible paths through the feature HMM, $P(r_0|q_t)$ the initial state probability of the feature HMM, $p(x_n^s|r_l, q_j)$ the probability of emitting feature component x_n^s while in feature HMM state r_l of temporal state q_j, and $P(r_l|r_{l-1}, q_j)$ the transition probabilities of the feature HMM in temporal state q_j.

We believe that HMM2 (which includes the classical mixture of Gaussian HMMs as a particular case) has several potential advantages, including:

1. Better feature correlation modeling through the feature HMM topology (e.g., working in the frequency domain). Also, the complexity of this topology and the probability density function associated with each state easily control the number of parameters.
2. Automatic non-linear spectral warping. In the same way the conventional HMM does time warping and time integration, the feature-based HMM performs frequency warping and frequency integration.
3. Dynamic formant trajectory modeling. As further discussed below, the HMM2 structure has the potential to extract some relevant formant structure information, which is often considered as important to robust speech recognition.

To illustrate these advantages and the relationship of HMM2 with dynamic multi-band ASR, we trained all parameters of an HMM2, using frequency filtered filterbank features [25]. We employed the HMM2 topology as shown in Figure 6. Training was done with the EM algorithm, and decoding was performed using the Viterbi algorithm for both the temporal and the frequency HMM. Figure 7 illustrates (on unseen test data) the temporal and frequency segmentation obtained as a by-product from Viterbi, plotted onto a spectrogram of our features. At each time step, we kept the 3 positions where the feature HMM changed its state during decoding (for instance, at the first time frame, the feature HMM goes from state r_1 to state r_2 after the second feature). We believe that this segmentation gives cues about some structures of the speech signal such as formant positions. In fact, in [31] it has been shown that this segmentation information can be used as (additional) features for speech recognition, being (1) discriminant and (2) rather robust in the case of speech degraded by additive noise.

7. Conclusions.

In this paper, we have discussed a family of new ASR approaches that have recently been shown to be more robust to noise, without requiring specific adaptation or "multi-style" training.

From all this discussion, and the convergence of independent experiments, we can draw the following preliminary conclusions:

1. Multiband ASR does not seem to be inherently inferior to a full-band approach, although some correlation information is lost due to the division of the frequency space into subbands.[19] Furthermore, it is not clear either that human hearing is using this kind of correlation information.
2. When training subband systems, we should not aim at maximizing the classification performance for every subband. When using the right combination rule, it should be better to increase the number of subbands while making sure that at any time at least one subband will be guessing the right answer.[20]
3. Doing this, we should also look at the potential for improvement in subband modeling when combining longer time-scale information streams (trading frequency information for temporal information).
4. The full combination approach discussed here has the potential of providing us with new adaptation schemes in which only the combination weights are automatically adapted (e.g., according to an online EM algorithm).

Finally, it is clear that several key problems remain to be addressed, including:

1. Need for improved expert weighting
2. Need for methods which are robust to noise but still perform well for clean speech.

In subband processing, there is also a need to properly choose the frequency subband, and it is expected that those subbands should be dynamically defined, e.g, following some formant structure. In this respect, the HMM2 formalism also presented here can be considered as a generalization of subband approaches, allowing for optimal (according to a maximum likelihood criterion) subband segmentation and recombination.

Acknowledgments. The content and themes discussed in this paper largely benefited from the collaboration with our colleagues Andrew Morris and Astrid Hagen. This work was partly supported by the Swiss Federal Office for Education and Science (FOES) through the European SPHEAR (TMR, Training and Mobility of Researchers) and RESPITE (ESPRIT Long term Research) projects. Additionally, Katrin Weber is supported by a the Swiss National Science Foundation project MULTICHAN.

[19] Probably because the advantages of subband based ASR can outweight the slight problem due to independent processing of subbands.

[20] This conclusion is very similar to what is proved mathematically in [4], p. 369, para. 1 (also p. 424).

REFERENCES

[1] Allen, J., “How do humans process and recognize speech?,” *IEEE Trans. on Speech and Audio Processing*, vol. 2, no. 4, pp. 567–577, 1994.

[2] Bengio, S., Bourlard, H. and Weber, K., “An EM Algorithm for HMMs with Emission Distributions Represented by HMMs,” *IDIAP Research Report*, IDIAP-RR-00-11, 2000.

[3] Berthommier, F. and Glotin, H., “A new SNR-feature mapping for robust multistream speech recognition,” *Intl. Conf. of Phonetic Sciences (ICPhS’99)* (San Francisco), to appear, August 1999.

[4] Bishop, C.M., *Neural Networks for Pattern Recognition*, Clarendon Press (Oxford), 1995.

[5] Bourlard, H. and Morgan, N., *Connectionist Speech Recognition – A Hybrid Approach*, Kluwer Academic Publishers, 1994.

[6] Bourlard, H. and Dupont, S., “A new ASR approach based on independent processing and combination of partial frequency bands,” *Proc. of Intl. Conf. on Spoken Language Processing* (Philadelphia), pp. 422-425, October 1996.

[7] de Veth, J., de Wet, F., Cranen, B., and Boves, L., “Missing feature theory in ASR: make sure you miss the right type of features,” *Proceedings of the ESCA Workshop on Robust Speech Recognition* (Tampere, Finland), May 25-26, 1999.

[8] Duda, R.O., Hart, P.E., *Pattern Classification and Scene Analysis*, John Wiley, 1973.

[9] Greenberg, S., “On the origins of speech intelligibility in the real world,” *Proc. of the ESCA Workshop on Robust Speech Recognition for Unknown Communication Channels*, pp. 23-32, ESCA, April 1997.

[10] Hagen, A., Morris, A., Bourlard, H., “Subband-based speech recognition in noisy conditions: The full combination approach,” *IDIAP Research Report no. IDIAP-RR-98-15*, 1998.

[11] Hagen, A., Morris, A., Bourlard, H., “Different weighting schemes in the full combination subbands approach for noise robust ASR,” *Proceedings of the Workshop on Robust Methods for Speech Recognition in Adverse Conditions* (Tampere, Finland), May 25-26, 1999.

[12] Hennebert, J., Ris, C., Bourlard, H., Renals, S., and Morgan, N. (1997), “Estimation of Global Posteriors and Forward-Backward Training of Hybrid Systems,” *Proceedings of EUROSPEECH’97* (Rhodes, Greece, Sep. 1997), pp. 1951-1954.

[13] Hermansky, H. and Morgan, N., “RASTA processing of speech,” *IEEE Trans. on Speech and Audio Processing*, vol. 2, no. 4, pp. 578-589, October 1994.

[14] Hermansky, H., Pavel, M., and Tribewala, S., “Towards ASR using partially corrupted speech,” *Proc. of Intl. Conf. on Spoken Language Processing* (Philadelphia), pp. 458-461, October 1996.

[15] Hermansky, H. and Sharma, S., “Temporal patterns (TRAPS) in ASR noisy speech,” *Proc. of the IEEE Intl. Conf. on Acoustics, Speech, and Signal Processing* (Phoenix, AZ), pp. 289-292, March 1999.

[16] Houtgast, T., Steeneken, H.J.M., “A review of the MTF concept in room acoustics and its use for estimating speech intelligibility in auditoria,” *J. Acoust. Soc. Am.*, vol. 77, no. 3, pp. 1069-1077, March 1985.

[17] S. Ikbal, H. Bourlard, S.Bengio, and K. Weber, “IDIAP HMM/HMM2 System: Theoretical Basis and Software Specifications” *IDIAP Research Report*, IDIAP-RR-01-27, 2001.

[18] Kingsbury, B., Morgan, N., and Greenberg, S., “Robust speech recognition using the modulation spectrogram,” *Speech Communication*, vol. 25, nos. 1-3, pp. 117-132, 1998.

[19] Lippmann, R.P., Carlson, B.A., “Using missing feature theory to actively select features for robust speech recognition with interruptions, filtering and noise,” *Proc. Eurospeech’97* (Rhodes, Greece, September 1997), pp. KN37-40.

[20] McGurk, H. and McDonald, J., “Hearing lips and seeing voices,” *Nature*, no. 264, pp.746-748, 1976.

[21] Mirghafori, N. and Morgan, N., “Transmissions and transitions: A study of two common assumptions in multi-band ASR,” *Intl. IEEE Conf. on Acoustics, Speech, and Signal Processing*, (Seattle, WA, May 1997), pp. 713-716.

[22] Morris, A.C., Cooke, M.P., and Green, P.D., “Some solutions to the missing features problem in data classification, with application to noise robust ASR,” *Proc. Intl. Conf on Acoustics, Speech, and Signal Processing*, pp. 737-740, 1998.

[23] Morris, A.C., Hagen, A., Bourlard, H., “The full combination subbands approach to noise robust HMM/ANN-based ASR,” *Proc. of Eurospeech’99* (Budapest, Sep. 99), to appear.

[24] Moore, B.C.J., *An Introduction to the Psychology of Hearing* (4th Edition), Academic Press, 1997.

[25] Nadeu, C., Hernando, J., and Gorricho, M., “On the decorrelation of filter-bank energies in speech recognition,” *Proc. of Eurospeech’95* (Madrid, Spain), pp. 1381-1384, 1995.

[26] Okawa, S., Bocchieri, E., Potamianos, A., “Multi-band speech recognition in noisy environment,” *Proc. IEEE Intl. Conf. on Acoustics, Speech, and Signal Processing*, 1998.

[27] Rao, S. and Pearlman, W.A., “Analysis of linear prediction, coding, and spectral estimation from subbands,” *IEEE Trans. on Information Theory*, vol. 42, pp. 1160–1178, July 1996.

[28] Tomlinson, .J., Russel, M.J., Brooke, N.M., “Integrating audio and visual information to provide highly robust speech recognition,” *Proc. of IEEE Intl. Conf. on Acoustics, Speech, and Signal Processing* (Atlanta), May 1996.

[29] Tomlinson, M.J., Russel, M.J., Moore, R.K., Bucklan, A.P., and Fawley, M.A., “Modelling asynchrony in speech using elementary single-signal decomposition,” *Proc. of IEEE Intl. Conf. on Acoustics, Speech, and Signal Processing* (Munich), pp. 1247-1250, April 1997.

[30] Varga, A. and Moore, R., “Hidden markov model decomposition of speech and noise,” *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing*, pp. 845–848, 1990.

[31] Weber, K., Bengio, S. and Bourlard, H., “HMM2- Extraction of Formant Features and their Use for Robust ASR”, *Proc. of Eurospeech*, pp. 607-610, 2001.

[32] Wellekens, C.J., Kangasharju, J., Milesi, C., “The use of meta-HMM in multistream HMM training for automatic speech recognition,” *Proc. of Intl. Conference on Spoken Language Processing* (Sydney), pp. 2991-2994, December 1998.

[33] Wu, S.-L., Kingsbury, B.E., Morgan, N., and Greenberg, S., “Performance improvements through combining phone and syllable-scale information in automatic speech recognition,” *Proc. Intl. Conf. on Spoken Language Processing* (Sydney), pp. 459-462, Dec. 1998.