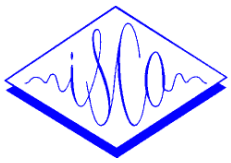

3rd International Conference on
Spoken Language Processing (ICSLP 94)
Yokohama, Japan
September 18-22, 1994

A PROBABILISTIC FRAMEWORK FOR WORD RECOGNITION USING PHONETIC FEATURES

Christoph Windheuser (*) *Frédéric Bimbot* (*) *Patrick Haffner* (**)

(*) Télécom Paris, Département Signal, CNRS, URA 820, Paris, France
(**) France Télécom, CNET LAA/TSS/RCP, Lannion, France

ABSTRACT

In this paper we propose a way to include a phonetic representation using binary distinctive phonetic features into the probabilistic framework of a hybrid system for word recognition. This could be done under two assumptions: 1.) The features are changing synchronously at the borders of the phonemes and 2.) the features are conditionally independent with one another. We report experiments we have done with a hybrid system working with a feature representation with which we get a recognition rate of 96.9 % word recognition on a spelled letter task. Then we describe some ways to weaken the assumption of independence between the features. We demonstrate that these ways do not improve significantly the recognition rate but lead to a simpler system with the same performance.

1. INTRODUCTION

A popular approach for an artificial speech recognition system is a *hybrid system* (see for example [1]). These systems consist of a connectionist component to estimate the *a posteriori* probabilities of a set of *states* (for example phonemes) and a non-connectionist time alignment procedure to find an optimal match between these estimated probabilities and an internal set of word models. A common choice for the connectionist part is a *Multilayer Perceptron* [2] or a *Time Delay Neural Network* (TDNN), [3] and the time alignment is usually done by a *Hidden Markov Model* [4] or by a *Dynamic Time Warping (DTW)* algorithm [5] which is based on the *Viterbi algorithm* [6].

A set of phonemes is usually chosen as the internal representation of the words in a hybrid system. The *a posteriori* probability of a phoneme given the acoustic input is estimated by the connectionist network. As we have reported previously [7], an alternative representation based on *binary discriminative phonetic features* can be used, which leads to a more compact system. In this paper we will show how the phonetic feature based representation can be included in the probabilistic framework of a hybrid system.

2. HYBRID SYSTEMS

The task of an artificial word recognition system is to recognize a set of words: $W = w_1, \ldots, w_N$. For a given sequence of acoustic vectors X_1^T, the system should determine the word w_l with the maximum *a posteriori* probability $P(w_l|X_1^T)$. Because this probability can not be computed directly, it is obtained by the Bayes rule:

$$P(w_l|X_1^T) = \frac{P(X_1^T|w_l)P(w_l)}{P(X_1^T)} \qquad (1)$$

$P(X_1^T|w_l)$ can be interpreted as the likelihood of the speech data given a *model* of the word w_l. $P(w_l)$ is the *a priori* probability of the word w_l, which can be computed by counting the number of occurences for every word. If we are just looking for the maximum value of $P(w_l|X_1^T)$, the denominator $P(X_1^T)$ can be ignored because it is common to all words w_l.

Every model of a word w_l is defined as a sequence of states $w_l = q_{l1}, \ldots, q_{ln}$. These states usually correspond to phonemes or parts of phonemes. In the alignment procedure every acoustic vector x_n of the input sequence X_1^T is assigned to a certain state q_{li} of a word w_l, which results in a sequence of states $Q_{l1}^T = q_{l1} \ldots q_{lT}$. To compute the likelihood $P(X_1^T|w_l)$, all possible alignments have to be taken into account, but the *Viterbi assumption* allows us to just use the alignment with the maximal likelihood. So we can write:

$$P(X_1^T|w_l) = \max_{Q_{l1}^T} P(X_1^T, Q_{l1}^T|w_l) \qquad (2)$$

and further:

$$P(X_1^T, Q_{l1}^T|w_l) = P(X_1^T|Q_{l1}^T) \cdot P(Q_{l1}^T|w_l) \qquad (3)$$

If we assume that all transition probabilities between the states are equal for every word, the second factor in formula (??) becomes a constant for a given speech input X_1^T. If we now make the usual conditional independence assumption, we get:

$$P(X_1^T|Q_{l1}^T) = \prod_{t=1}^{T} P(x_t|q_{lt}) \qquad (4)$$

so that we can write:

$$P(X_1^T|w_l) \propto \max_{q_{lt} \in Q_{l1}^T} \prod_{t=1}^{T} P(x_t|q_{lt}) \qquad (5)$$

This means that the likelihood of an utterance given a word model is proportional to the accumulated multiplications of the likelihood of a frame given a state along the optimal path found by the alignment procedure.

As it has been shown for example in [8, 9], a *Multilayer Perceptron* trained with a least mean square error or a conditional entropy error function approximates the *a posteriori* probability of a class given the input without any specific assumption about the distribution of the data. This fact is used in a hybrid system. Here the connectionist part estimates the *a posteriori* probability of a state given the speech input $P(q_i|x_t)$. To get the likelihood $P(x_t|q_i)$, which is needed in formula 5, we once again apply the Bayes rule.

3. PHONETIC FEATURES

Binary distinctive phonetic features were introduced as a system to describe phonemes and phones according to their similarities and opositions [10]. They can be viewed as a general representation of the phonemes. Converserly, any binary representation of the phonemes can be viewed as a feature representation. Note that the traditional 1-out-of-N representation is a special (canonical) case of a feature representation (when every phoneme is represented by one feature). Previous work has shown that Multilayer Perceptrons trained to recognize phonetic features can gain high recognition scores [11, 12, 13, 14], which makes them suitable for speech recognition.

For using phonetic features in a speech recognition task, a set of features has to be defined. As a necessary condition it must be possible to distinguish the words or the phonemes by their phonetic feature representation, i.e.: the features have to be *discriminant*. Furthermore previous experiments have illustrated the fact that features which rely mainly on acoustic or articulatory properties are easier to detect in the acoustic signal than arbitrary features.

In [7] we have reported several experiments with different feature sets. We got our best results on a spelling letter task with a set consisting of 18 features (see table 1). In this set we have 7 features for the vowels (*vowel, closed vowel, mid vowel, open vowel, back vowel, central vowel* and *front vowel*), 9 features for the consonants (*vocalic consonant, liquid, nasal, consonant, plosive, fricative, front consonant, central consonant* and *back consonant*) and two general features (*unvoiced* and *voiced*). Although some features are redundant with others, we originally used this system because every phoneme was characterized by 4 active and 14 inactive features.

4. PHONETIC FEATURES IN A HYBRID SYSTEM

To include a phonetic feature representation into a hybrid system, we represent every phoneme by its phonetic feature representation. This implies the assumption of a synchronous change of the features at the borders of the phonemes (for a discussion see for example [15]).

We start with a set of phonetic features $f_1 \dots f_i \dots f_M$ (M is 18 in our case). For every phoneme p_j some features are set to 1 *(active)* and the others are set to 0 *(inactive)*:

$$\begin{aligned} p_j &= (f_1 = b_{j1}, \dots, f_i = b_{ji}, \dots, f_M = b_{jM}) \quad (6) \\ b_{ji} &= 0 \vee 1 \end{aligned}$$

In our second assumption we say that all features are conditionally independent and each feature combination corresponds to a phoneme. With this assumption we can express the *a posteriori* probability of a phoneme given the input $P(p_j|x_t)$ by the *a posteriori* probability that a feature is active given the input $P(f_i = 1|x_t)$:

$$\begin{aligned} P(p_j|x_t) &= \prod_{i=1}^{M} P(f_i = b_{ji}|x_t) \quad (7) \\ &= \prod_{b_{ji}=1} P(f_i = 1|x_t) \prod_{b_{ji}=0} (1 - P(f_i = 1|x_t)) \end{aligned}$$

$P(f_i = 1|x_t)$ is estimated by a multilayer perceptron which is trained on feature recognition. If we replace the *a posteriori* probabilities $P(p_j|x_t)$ by the likelihood $P(x_t|p_j)$ using the Bayes rule and skip the probability $P(x_t)$, we can include formula (7) into the framework of a hybrid system:

$$P(X_1^T|w_l) \propto \quad (8)$$

$$\prod_{\substack{t=1 \\ p_j \in \text{best path}}}^{T} \frac{1}{P(p_j)} \prod_{b_{ji}=1} P(f_i = 1|x_t) \prod_{b_{ji}=0} (1 - P(f_i = 1|x_t))$$

$P(p_j)$ being the *a priori* probabilities of the phonemes.

5. EXPERIMENTS

In our hybrid system we use a TDNN to estimate the *a posteriori* probabilities $P(f_i = 1|x_t)$. The network consists of an input layer with 16 units for the acoustic vector, a hidden layer with 24 units and a output layer with 18 units for the features. The weights between the input and the hidden layer have the delays -1, 0 and 1. The weights between the hidden layer and the output layer have the delays -2, 1, 0, 1 and 2. Together with the bias this adds up to a total number of 3354 learnable parameters of the network, which is quite small for a hybrid speech recognition system.

For the training we use the plain on-line backpropagation algorithm with a mean square error function at the frame level. The feature labels are provided by a HMM system working in forced-alignment mode. The learning rate is set to 1.0.

5.1. The Task

We trained our system on a speaker dependent word recognition task consiting of the spelling letters of the english alphabet. Despite the small vocabulary (26 "words"), this is a difficult task because a lot of words are highly confusible. We used the ALPH database from CMU (speaker jmt). This database contains 1000 sequences of spelled letters. In every sequence, on the average 5 letters are spelled continuously. The speech was recorded in high quality and transformed into a bark-scale representation of 16 coefficients every 10 ms. We split the base into 500 sequences for training and 500 sequences for testing.

After a training of 150 epochs on the training set, we got a word recognition rate of 96.9 % on the test set and 97.0 % on the training set. The small difference between the test and the training set illustrates a good generalisation abillity of the network, which is due to the small number of parameters.

6. THE INDEPENDENCE ASSUMPTION OF THE FEATURES

The assumption of a conditional independence of the features is obviously not true in practice. Features like *voiced* and *unvoiced* for example are totally correlated. A mathematical modeling of the dependencies is not straighforward, because the correlations between the features are conditioned by the speech input x_t. To circumvent this problem, we can try to make the features less dependent.

6.1. Discarding Redundant Features

Our feature set was designed with a lot of redundancy. This was done with the idea in mind that the redundancy makes the speech recognition more robust. On the other hand, the redundancy creates correlations between the features. To reduce the redundancy, we suppressed the features *vowel, central vowel, consonant, vocalic consonant* and *voiced*, which are totally predicteable from the other

features. In practice we used only 13 of the previous 18 outputs of the network, which we combined according to equation 8.

Interestingly, with this approach we get the same result of 96.9 % word recognition rate on the test set. So we did not improve the results, but we get the same results with a significantly smaller number of features.

6.2. Grouping the Features

Another possibility to reduce the dependencies between the features is to define *groups of features* $g_1 \dots g_i \dots g_N$. Every group consists of a number of different features: $g_i = (f_{i1}, \dots, f_{ik})$. The features inside a group have mutually exclusive outcomes, i.e. exactly one feature of a group is active and all other features are inactive. Now the representation of a phoneme is defined by specifying for every group, which feature is active:

$$\begin{aligned} p_j &= (g_1 = c_{j1}, \dots, g_i = c_{ji}, \dots, g_N = c_{jN}) \quad (9) \\ c_{ji} &\in \{i_1, \dots, i_k\} \\ c_{ji} &= i_l \implies f_{il} = 1 \end{aligned}$$

Now instead of equation 7, the *a posteriori* probability of the phoneme p_j can be expressed by:

$$P(p_j|x_t) = \prod_{i=1}^{N} P(g_i = c_{ji}|x_t) \quad (10)$$

Here, we assume the conditional independence between every group of features. Because inside every group we have an 1-out-of-N representation, we can set:

$$P(g_i = c_{ji}|x_t) = P(f_{c_{ij}} = 1|x_t) \quad (11)$$

We have divided our set of features into 4 groups ($N = 4$). The first group describes the general role of the phoneme *(vowel, consonant, vocalic consonant)*, the second group describes the place of articulation *(back vowel, central vowel, front vowel, back consonant, central consonant, front consonant)*, the third group describes the mode of articulation *(liquid, nasal, plosive, fricative, closed vowel, mid vowel, open vowel)* and the last group describes the viocingof the phoneme *(voiced, unvoiced)*.

By testing this approach, we got again the same results as with the original system (96.9 % word recognition rate on the test set), although we again have a simpler representation of the features.

6.3. Minimal Number of Features

As we have seen, we can reduce the number of features without deteriorating the results. To make the phoneme representation even more compact, we just take into account the minimal number of features which are necessary to distinguish this phoneme from the others. To identify the phoneme */m/* for example, it is sufficient to know that the feature *nasal* is active to distinguish it from all other non nasal phonemes and that the feature *front consonant* is active to distinguish it from the other nasal phoneme */n/*. If we do this systematically for all phonemes, we find that two or three features per phoneme are sufficient to distinguish each of them from the other phonemes.

By taking just this minimal number of features per phoneme into account, we could slightly (but not significantly) improve the results to 97.0 % word recognition rate on the test set.

6.4. Modeling the Dependencies

Without the assumption of the independence of the features, formula (7) becomes:

$$\begin{aligned} P(p_j|x_t) &= P(f_1 = b_{j1}, \dots, f_M = b_{jM}|x_t) \quad (12) \\ &= P(f_1 = b_{j1}|x_t) \cdot P(f_2 = b_{j2}|f_1 = b_{j1}, x_t) \cdots \\ &\quad P(f_M = b_{jM}|f_1 = b_{j1}, \dots, f_{M-1} = b_{jM-1}, x_t) \end{aligned}$$

All the factors of formula (12) could be estimated by a neural network. This would result in a modular hierarchical architecture of networks. The first network would estimate the *a posteriori* probability of feature f_1 given the observation x_t. f_1 would be the most general feature (for example *voiced*). The next network would take as input the observation x_t and the target value for feature f_1 i.e. b_{j1}. The next network would take as input x_t, b_{j1} and b_{j2}. This continues until network M, which would have $M-1$ additional inputs to indicate the states of all other features.

Another possibility would be to model the dependencies between the features by a function f:

$$\begin{aligned} P(p_j|x_t) &= f(P(f_1 = b_{j1}|x_t), P(f_2 = b_{j2}|x_t), \\ &\quad \dots, P(f_M = b_{jM}|x_t)) \quad (13) \end{aligned}$$

If we constrain function f to be a member of a given family of functions and characterized by a set of parameters, these parameters could be estimated from the training set.

These two approaches are being investigated at the moment, therefore we can not report any results yet.

7. CONCLUSION

We have presented ways to include a binary distinctive phonetic feature representation of the phonemes into a probabilistic framework of a hybrid system. We had to make two assumptions:

- The features change synchronously at the borders of the phonemes.
- The features are conditionally independent.

With the first assumption we could use the dynamic programming algorithm and the second assumption allows us to easily compute the *a posteriori* probability of a phoneme given the *a posteriori* probabilities of the features.

Despite the fact that these two assumptions are not true in practice, we obtained high recognition results on our task with a simple and compact system. In a second step we showed several ways to weaken the independence assumption of the features.

The fact that reducing the dependency between the features did not improve the recognition results tends to show that the independence assumption is not too critical here. However a more elaborate way of dealing with the dependencies may further improve the results.

8. ACKNOWLEDGEMENTS

The authors would like to thank Alex Waibel for providing the ALPH database. They gratefully acknowledge financial support from the French *Ministère de l'Enseignement Supérieur et de la Recherche* and from *France Télécom, CNET Lannion.*

Table 1. Definition of the phonetic feature set. Because in our task it was not necessary to distinguish between the phonemes */y/* and */iy/* as well as between the phonemes */w/* and */uw/*, these two pairs have the same feature representation. Furthermore we split the phonemes */ey/* into */eh/* and */y/*, */ay/* into */aa/* and */y/*, */ch/* into */t/* and */ch/*, */jh/* into */d/* and */jh/* and */ow/* into */ow/* and */w/*.

	aa	ah	ax	eh	y	iy	ow	uw	l	m	n	jh	r	z	v	w	d	b	s	ch	f	t	k	p
vowel	+	+	+	+	+	+	+	+	−	−	−	−	−	−	−	+	−	−	−	−	−	−	−	−
closed vowel	−	−	−	−	+	+	−	+	−	−	−	−	−	−	−	+	−	−	−	−	−	−	−	−
mid vowel	−	+	+	+	−	−	+	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−
open vowel	+	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−
back vowel	−	−	+	−	−	−	+	+	−	−	−	−	−	−	−	+	−	−	−	−	−	−	−	−
central vowel	+	+	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−
front vowel	−	−	−	+	+	+	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−
vocalic cons.	−	−	−	−	−	−	−	−	+	+	+	−	+	−	−	−	−	−	−	−	−	−	−	−
liquid	−	−	−	−	−	−	−	−	+	−	−	−	+	−	−	−	−	−	−	−	−	−	−	−
nasal	−	−	−	−	−	−	−	−	−	+	+	−	−	−	−	−	−	−	−	−	−	−	−	−
consonant	−	−	−	−	−	−	−	−	−	−	−	+	−	+	+	−	+	+	+	+	+	+	+	+
plosive	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	+	+	−	−	−	+	+	+
fricative	−	−	−	−	−	−	−	−	−	−	−	+	−	+	+	−	−	−	+	+	+	−	−	−
front cons.	−	−	−	−	−	−	−	−	−	+	−	−	−	−	+	−	−	+	−	−	+	−	−	+
central cons.	−	−	−	−	−	−	−	−	+	−	+	−	−	+	−	−	+	−	+	−	−	+	−	−
back cons.	−	−	−	−	−	−	−	−	−	−	−	+	+	−	−	−	−	−	−	+	−	−	+	−
unvoiced	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	+	+	+	+	+	+
voiced	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	−	−	−	−	−	−

REFERENCES

[1] Bourlard, H.A. and Morgan, N., *Connectionist Speech Recognition*, Kluwer Academic Publishers, 1994.

[2] Rumelhart, D.E., Hinton, G.E. and Williams, R.J., *Parallel Distributed Processing*, (MIT Press, 1986), pp. 318-362.

[3] Waibel, A., Hanazawa, T., Hinton, G., Shikano, K. and Lang, K., "Phoneme Recognition using Time-Delay Neural Networks," *IEEE Transactions on Acoustics, Speech and Signal Processing*, Vol. 37, pp. 328-339, 1989.

[4] Jelinek, F., "Continuous Speech Recognition by Statistical Methods", *Proceedings of the IEEE*, Vol. 64, No.4, pp. 532-556, 1976.

[5] Sakoe, H. and Chiba, S., "Dynamic Programming Algorithm Optimization for Spoken Word Recognition," in: *Readings in Speech Recognition*, Morgan Kaufmann, pp. 159-165, 1990.

[6] Viterbi, A.J., "Error Bounds for Convolutional Codes and an Asymptotically Optimum Decoding Algorithm", *IEEE Trans. on Information Theory*, Vol. 13, No. 2, pp. 260-269, 1967.

[7] Windheuser, C. and Bimbot, F.: "Phonetic Features for Spelled Letter Recognition with a Time Delay Neural Network", *Proc. of Eurospeech*, pp. 1489-1492, Berlin, 1993.

[8] Bourlard, H.A. and Wellekens, C.J., "Links between Markov Models and Multilayer Perceptrons", *Computer, Speech and Language*, Vol. 3, pp. 1-19, 1989.

[9] Richard, M.D. and Lippmann, R.P., "Neural Network Classifiers Estimate Bayesian *a posteriori* Probabilities", *Neural Computation*, Vol. 3, pp. 461-483, 1991.

[10] Jakobson, R., Fant, C.G.M., and Halle, M., *Preliminaries to Speech Analysis: The Distinctive Features and their Correlates*, MIT Press, 1951.

[11] Bimbot, F., Chollet, G., and Tubach, J.-P. "TDNNs for Phonetic Feature Extraction: A Visual Exploration," *Proc. of the ICASSP*, 1991.

[12] De Mori, R. and Flammia, G. "Speaker-independent consonant classification in continuous speech with distinctive features and neural networks", in *J. Acoust. Soc. Am.*, Vol. 94, No. 6, pp. 3091-3103, 1993.

[13] Bradshaw, G. and Bell, A., "Towards the Performance Limits of Connectionist Feature Detectors", in *International Conference on Spoken Language Processing*, Banff, 1992.

[14] Okawa, S., Windheuser, C., Bimbot, F., and Shirai, K., "Evaluation of Phonetic Feature Recognition with a Time-Delay Neural Network", *Proc. of the International Conference on Spoken Language Processing ICSLP*, Yokohama, 1994.

[15] Huckvale, M., "The Benefits of Tiered Segmentation for the Recognition of Phonetic Properties", *Proc. of Eurospeech*, pp. 1473-1476, Berlin, 1993.