Speech Recognition with Phonological Features: Some issues to attend

Frederik Stouten and Jean-Pierre Martens

ELIS-Ghent University
Sint-Pietersnieuwstraat 41, B-9000 Ghent, Belgium
{fstouten,martens}@elis.ugent.be

Abstract

It is often argued that acoustic-phonetic or articulatory features could be beneficial to automatic speech recognition because they provide a convenient interface between the acoustic and the linguistic level. Former research has shown that a combination of acoustic and articulatory information can lead to improved ASR. However there exists no purely articulatory driven ASR system that outperforms state-of-the-art systems driven by acoustic features. In this paper we propose a novel method for improving ASR on the basis of articulatory features. It is designed to take account of (1) the correlations between articulatory features and (2) the fact that not all articulatory features are relevant for the description of a certain phonetic unit. We also investigate to what extend an acoustic and an articulatory feature driven system make different errors.
Index Terms: speech recognition, phonological features, decorrelation, relevancy

1. Introduction

The use of phonetic or articulatory features for ASR has been studied for more than a decade. However, the term articulatory features covers a variety of concepts, from phonological features (PHFs) used in phonological sound categorization (e.g. [1]) to acoustic properties that are presumed to correlate with articulatory measurements [2]. In this paper we deal with PHFs. The main reasons for using such features are,

- PHFs constitute an intermediate level between MFCCs and phonemes, perhaps the highest information level that can be extracted reliably from the speech signal.
- One PHF is involved in the characterization of multiple phones and multiple languages. Training material can thus be shared across phones and languages, which may offer a basis for better multilingual and cross-lingual ASR.
- Pronunciation variations can be naturally described in terms of phonological feature overlap and assimilation [3].

The need for separate stochastic models to extract the PHFs adds complexity to the recognition system, and the question is of course whether this additional effort is justified. Most of the work on PHF-based ASR has focused on phoneme recognition ([4, 5, 6]) or on small vocabulary word recognition ([7]). Nevertheless, it has been shown ([8, 9, 10]) that combining PHFs and MFCCs can raise the performance of large vocabulary continuous speech recognition (LVCSR) systems. In Metze et al. [9] adding 6 to 10 well chosen PHFs to the MFCC stream yields a 15 % relative reduction of the WER for a read BN task, and a 7.5 % reduction on a spontaneous scheduling task. In her PhD, Kirchhoff [10] investigated several state-level and word-level combination techniques. With a state-level combination technique the WER dropped from 29.03 % to 27.41 % (German Verbmobil corpus). With a word-level combination technique a WER of 27.97 % WER was obtained. The purely PHF-driven system however performed worse than the acoustic baseline. We owe this, for a part, to the suboptimal use of PHFs in the standard HMM framework which is after all optimized for MFCCs. In this paper we investigate how to adapt this framework so as to raise the accuracy of purely PHF-based ASR. With a higher performance we also hope to achieve larger gains by combining a PHF and a MFCC-based system.

The rest of the paper is organized as follows. In Section 2 we provide details about the PHFs we are using and the way they are extracted. In Section 3 we discuss two major problems with respect to the application of PHFs in the standard HMM framework and we outline possible solutions to these problems. In Section 4 we propose an integrated technique to implement these solutions, and in Section 5 we present an experimental evaluation.

2. The Phonological Features

In a former paper [11] we proposed a set of PHFs that is presumed to meet the following two criteria: (1) on the basis of phonological knowledge, it is easy to attribute canonical feature values to all the phonetic units, and (2) it is possible to extract the features in a reliable way by means of an automatically trained system. The chosen feature set consists of 27 binary features which are organized along 4 feature dimensions:

- **vocal source**: voiced, unvoiced, no-activation
- **manner**: closure, vowel, fricative, burst, nasal, approximant, lateral, silence
- **place-consonant**: labial, labio-dental, dental, alveolar, post-alveolar, velar, glottal
- **vowel-features**: low, mid-low, mid-high, high, back, mid, front, retroflex, round

The vocal source describes the frame-level presence/absence of speech and the nature (voiced/ unvoiced) of that speech. The other features of a frame describe properties of the phonetic unit to whose realization that frame is contributing. Their detection requires inputs from a larger time interval. For instance, the distinction between a *closure* and a *silence* resides in the length of the no-activation interval.The features are detected by means of a hierarchical system comprising four multi-layer perceptrons (MLPs) (see Figure 1 and [11]). We only used the *voiced* output from the vocal source network, since the two other features were largely covered by the other networks. Important is that the PHFs are computed starting from the MFCCs, that the MLPs are supplied with a sequence of subsequent MFCC vectors, and that MLP outputs are presumed to represent posterior probabilities of the binary PHFs. During training, vowel frames for instance do not contribute to the training of the place consonant MLP. This means that place consonant features are considered irrelevant for the description of vowel

frames, and they should not force the consonant place MLP to produce prescribed outputs (e.g. zero) for these frames.

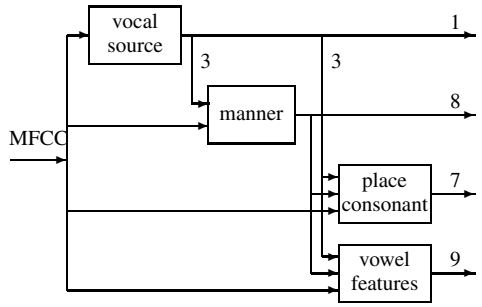

Figure 1: The phonological feature extractor.

On TIMIT data, the accuracy of the *manner* MLP was 83.9 %, that of the place-consonant MLP was 83.2 %. These figures are in line with those reported in [10, 12].

3. Properties of MFCCs and PHFs

In this section we discuss two important differences between MFCCs and PHFs that call for an adaptation of the standard HMM framework to make it more effective in combination with PHFs.

3.1. Feature correlations

One of the interesting properties of MFCCs is that their components are largely uncorrelated. This means that state-level emission distributions can take the form of Gaussian mixture models (GMMs) with only a few diagonal covariance mixtures. The binary PHFs on the other hand are expected to exhibit much larger correlations. The consonant place of articulation for instance is represented by 7 binary features, which will inevitably be correlated. Consequently the required emission distributions may no longer be represented efficiently by diagonal covariance GMMs. One solution would be to replace them by full covariance GMMs, but this would severely increase the number of model parameters per Gaussian. It may be more efficient to adopt one of the following techniques instead:

- Feature selection that aims at removing features carrying information largely carried by other features.
- Global decorrelation schemes such as PCA to transform the feature space in the front-end.
- Decorrelation on the basis of state-dependent feature transformation matrices.

Since one PHF is not consistently (across states) correlated with another, the first technique does not seem to be an option. Since the correlations are bound to depend very much on the state (e.g. for a particular consonant mainly the correlations between 2 or 3 of the near-canonical place features will be important to model), a global decorrelation scheme is not considered a viable option either. We therefore explored the third technique.

Obviously, transforming the features in a given state and modeling the transformed features with diagonal covariance GMMs is equivalent to modeling the non-transformed features with full covariance GMMs. Gales [13] has elaborated this approach and proposed a ML methodology to simultaneously train state-dependent linear transformation matrices (MLLT-matrices) and emission distributions. We will adopt Gales' method and extend it in a way that it can also deal with another problem that is typical for PHFs and that is discussed in the next subsection.

3.2. Feature relevancy

From the description of our PHFs and their training (see also [11]) it follows that not all features are relevant for all phones. Consequently, the emission distributions on a particular state should only be modeled in the subspace of the relevant features for that state. However, since working with different subspaces on different states causes problems with respect to the equivalences of likelihoods, the observation likelihoods may need to be factorized as the product of a relevant and an irrelevant observation likelihood. The latter can then be computed on the basis of a global model by adopting principles of missing data theory [14], more in particular by likelihood imputation. In the next section we show that likelihood imputation and state-dependent feature transformations can be embedded in a consistent probabilistic framework.

4. The proposed probabilistic framework

In order to formulate the re-estimation formulas we introduce m as an index pointing to one of the Gaussians to model. A particular Gaussian can be used in different states, but all these states have the same sets of relevant and irrelevant features respectively.

If we assume that the global model is a multivariate Gaussian distribution with a full covariance matrix and if $S_R(i)$ is the set of Gaussians for which feature i is relevant and $S_I(i)$ the complementary set, then the global model parameters (μ^g, Σ^g) can be updated according to

$$(\mu^g)_i = \frac{\sum_{m\in S_I(i)} \sum_{t=1}^{T} \zeta_t(m)(\mathbf{x}_t)_i}{\sum_{m\in S_I(i)} \sum_{t=1}^{T} \zeta_t(m)} \tag{1}$$

$$(\Sigma^g)_{ij} = \frac{\sum_{m\in S_I(i)\cap S_I(j)} \sum_{t=1}^{T} \zeta_t(m)(\mathbf{x}_t-\mu^g)_i(\mathbf{x}_t-\mu^g)_j}{\sum_{m\in S_I(i)\cap S_I(j)} \sum_{t=1}^{T} \zeta_t(m)} \tag{2}$$

with $\zeta_t(m)$ representing the probability of being in component m at time t and with $\mathbf{x}_t$ being the feature vector at time t.

If $\mathrm{R}(m)$ and $\mathrm{I}(m)$ represent the relevant and irrelevant feature sets of Gaussian m, the means of the relevant features of that Gaussian can be updated as follows:

$$(\mu_m)_i = \frac{\sum_{t=1}^{T} \zeta_t(m)(\mathbf{x}_t)_i}{\sum_{t=1}^{T} \zeta_t(m)} \quad \forall i \in \mathrm{R}(m) \tag{3}$$

To update the covariance matrix, we define the so-called accumulator W_m as

$$(W_m)_{ij} = \begin{cases} \frac{\sum_{t=1}^{T} \zeta_t(m)(\mathbf{x}_t-\mu_m)_i(\mathbf{x}_t-\mu_m)_j}{\sum_{t=1}^{T} \zeta_t(m)} & \forall i,j \in \mathrm{R}(m) \\ \frac{\sum_{t=1}^{T} \zeta_t(m)(\mathbf{x}_t-\mu^g)_i(\mathbf{x}_t-\mu^g)_j}{\sum_{t=1}^{T} \zeta_t(m)} & \forall i,j \in \mathrm{I}(m) \\ 0 & \text{else} \end{cases}$$

with the zero indicating that no correlation between relevant and irrelevant features is being modeled. To obtain the covariance matrix update formula one must take the MLLT matrices into account. If there are P such matrices and if S_p is the set of Gaussians using the same matrix A_p then it can be shown (see [13]) that

$$\Sigma_m^R = \mathrm{diag}(A_p W_m A_p^t) \;\; \forall m \in S_p \tag{4}$$

The next step is then to update A_p. If feature i is irrelevant in S_p, then the i-th row/column of A_p is equal to the corresponding row/column of the unity matrix. If $\mathbf{x}_t^R$ and $\mathbf{x}_t^I$ comprise the relevant and the irrelevant components of observation vector $\mathbf{x}_t$ respectively,

the likelihood $b_m^{A_p}(\mathbf{x}_t)$ generated by Gaussian $m \in S_p$ can be factorized as

$$\frac{|A_p|}{(2\pi)^{d-q/2}|\Sigma_m^R|^{1/2}} e^{[-\frac{1}{2}[A_p^R(\mathbf{x}_t^R-\mu_m^R)]^t(\Sigma_m^R)^{-1}A_p^R(\mathbf{x}_t^R-\mu_m^R)]}$$
$$\frac{1}{(2\pi)^{q/2}|\Sigma^{g,I}|^{1/2}} e^{[-\frac{1}{2}(\mathbf{x}_t^I-\mu^{g,I})(\Sigma^{g,I})^{-1}(\mathbf{x}_t^I-\mu^{g,I})]}$$

which is the product of an informative part, modeled by Gaussian m and a part imputed from the global model:

$$b_m^{A_p}(\mathbf{x}_t) = b_m^{A_p}(\mathbf{x}_t^R|m)\; l_{imput}(\mathbf{x}_t^I|\mu^g, \Sigma^g) \quad (5)$$

We initialize all MLLT-matrices to the unity matrix and update the elements corresponding to relevant features using Gales' algorithm. It makes use of auxiliary matrices $G^{(k)}$ defined in terms of the previously defined accumulator W_m,

$$(G^{(k)})_{ij} = \sum_{m\in S_p} \frac{(W_m)_{ij}}{(\sigma_m^2)_k} \sum_{t=1}^{T} \zeta_t(m) \quad \forall k \in R(S_p) \quad (6)$$

Only the accumulators differ from those in [13]. If we denote p_k as the k-th row of the transposed adjunct matrix of the actual A_p [1], the k-th row of the new A_p is obtained as

$$(\hat{A}_p)_{ki} = \begin{cases} \alpha_k\, p_k[(G^{(k)})^{-1}]_i & \forall k,i \in R(S_p) \\ \delta_{ki} & \text{else} \end{cases} \quad (7)$$

with α being equal to

$$\alpha_k = \sqrt{\frac{\sum_{m\in S_p}\zeta_t(m)}{p_k(G^{(k)})^{-1}p_k^t}} \quad \forall k \in R(S_p) \quad (8)$$

The entire 4-step algorithm to estimate all model parameters, can be summarized as follows:

1. Re-estimate the global model by means of (1) and (2).
2. Re-estimate the Gaussians using (3) and (4).
3. Re-estimate the MLLT-matrices A_p using (7) and (8)
4. Got to 2 until convergence, or appropriate criterion satisfied

Note that step 3 of this algorithm is in itself an iterative step since the re-estimation formula for row k depends on the cofactor row p_k which is a function of all rows but row k. So, we need to iterate in order to obtain good estimates for the new rows.

5. Recognition Experiments

In this section we present results with two types of recognizers: MFCC- and PHF driven systems both with and without the application of the decorrelation/irrelevancy handling technique. All recognition experiments were performed with the HTK-toolkit [15]. The database is TIMIT, and the core test set (24 speakers × 8 sentences) is the test database. The LM is a back-off bigram learned from the training and test utterances (see [3]). The acoustic models are cross-word triphone HMMs with underlying tied distributions (GMMs). State tying was performed using DT-based clustering. The training involved 2 Baum-Welch re-estimation steps for each number of mixtures, and the number of mixtures was changed from 1 to 6. When applying decorrelation, we performed 10 iterations of our 4-step algorithm and we allowed 100 iterations in step 3 of that algorithm. Next, three more Baum-Welch iterations were performed.

[1] The adjunct matrix is obtained by replacing all elements of the matrix by their cofactors and by transposing this cofactor matrix

5.1. PHF-driven ASR

For the experiments with the PHF-driven system we used 25 features (from the vocal source features only *voiced* was retained) and their time derivatives (= 50 features per frame). Without the application of our decorrelation strategy we obtained the baseline results listed in Table 1. Using a block diagonal MLLT-matrix (two blocks of 25×25) per phoneme (the central phoneme of the triphone) and assuming that all features are relevant, we obtained the 41-MLLT results of Table 1. Apparently, the WER drops from 8.34 % to 6.11 % (improvement of 26 % relative) by applying decorrelation. By including irrelevancy modeling on top of that, the WER did not decrease any further. The best WER was 6.37 % now (see Table 2). A possible reason for this may be that the general model was too simple [2].

system	#mix	# pars.	WER	D	S	I
baseline	1	184300	10.51	14	122	29
41-MLLT	1	(+51250)	7.58	17	90	12
baseline	2	368600	9.43	16	105	27
41-MLLT	2	(+51250)	7.20	20	77	16
baseline	3	552900	9.17	14	103	27
41-MLLT	3	(+51250)	6.11	17	65	14
baseline	4	737200	8.34	15	90	26
41-MLLT	4	(+51250)	6.69	22	67	16
baseline	6	1105800	8.85	20	88	31
41-MLLT	6	(+51250)	6.18	15	66	16

Table 1: *WER (%) for the baseline PHF system (25 features+25 delta's) and the system with MLLT-matrices trained on all features, for different numbers of Gaussian components.*

41-MLLT with relevancy handling					
#mix	# pars.	WER	D	S	I
1	122278 (+23540)	9.24	23	104	18
2	244506 (+23540)	8.09	20	87	20
3	366734 (+23540)	7.58	19	83	17
4	488962 (+23540)	6.43	16	70	15
5	611190 (+23540)	6.37	18	66	16
6	733418 (+23540)	6.50	17	67	18

Table 2: *WER (%) for the MLLT system with relevancy handling for different numbers of Gaussian components*

5.2. MFCC-driven ASR

In order to asses the obtained PHF-based results we have also constructed a standard MFCC-based system with 39 input features. The performances of that system with and without decorrelation (all MFCC parameters are presumed to be relevant at all times) are listed in Table 3. As before we used 41 block-diagonal (3 blocks of 13 rows each now) MLLT matrices. Apparently, the decorrelation technique is effective here too: the WER drops from 4.59 % to 3.69 % (20 % relative improvement). What is also clear is that the MFCC system outperforms the PHF system.

[2] We used a diagonal covariance gaussian distribution, but we will test more complex models soon

system	#mix	# pars.	WER	D	S	I
baseline	1	99138	7.07	27	69	15
41-MLLT	1	(+20787)	6.43	22	64	15
baseline	2	198276	6.05	21	66	8
41-MLLT	2	(+20787)	5.29	25	50	8
baseline	3	297414	5.99	22	64	8
41-MLLT	3	(+20787)	4.46	18	44	8
baseline	4	396552	5.16	20	56	5
41-MLLT	4	(+20787)	4.20	16	43	7
baseline	6	594828	4.59	18	50	4
41-MLLT	6	(+20787)	3.69	17	36	5

Table 3: *WER (%) for the baseline MFCC system (39 parameters) and the MFCC system with MLLT-matrices for different numbers of Gaussian components.*

5.3. Combining the two systems

If we can show that the MFCC and PHF-driven systems behave differently, then we have an argument for investigating whether a combination of the systems would lead to a further improvement of the ASR performance. In order to show this, we have compared the errors made by the two-systems (best configuration for each) and we found (see Table 4) that for 5.3 % of the words, the MFCC and the PHF-based ASR generated a different result. If we would be able to correct all the errors of the MFCC system that correspond to a correct solution in the PHF system, and if we would be able to avoid the introduction of new errors at other places, the WER could be reduced from 3.7 % to 2.6 % (relative improvement = 30 %). Obviously we will not be able to conceive such a good combination strategy. On the other hand, the maximum attainable improvement may be larger if not only the top-1 hypotheses but the top-N hypotheses of the individual ASR systems were taken into account.

error type	count	(%)
both correct	1470	93.6
MFCC wrong and PHF correct	18	1.15
MFCC correct and PHF wrong	47	3.00
both wrong, different errors	18	1.15
both wrong, same errors	17	1.10
total #words	1570	100

Table 4: *Number of word errors in the outputs of MFCC and PHF recognizers.*

6. Conclusions

In this paper we have investigated how to adopt the standard HMM approach to ASR so as to achieve the best possible performance when no MFCCs but Phonological Features (PHFs) are used as the acoustic observations. It was argued that unlike MFCCs PHFs are strongly correlated, and moreover, not all PHFs are relevant in all states. On the basis of these arguments we have proposed a novel decorrelation and irrelevancy handling technique which can be considered as an extension of a decorrelation technique originally proposed by Gales [13]. Experiments on TIMIT have revealed that the WER of a PHF-based ASR can be reduced by about 26 % relative, but that this improvement is insufficient to bring the WER at the same level of that of a state-of-the art MFCC-based ASR system. Another finding is that the decorrelation approach is also very helpfull to improve the MFCC-based ASR (relative improvement of 20 %). Comparing the errors made by the best MFCC and PHF-based systems we were also able to establish that the two systems behave differently, and that there is support for the thesis that combining the hypotheses of both systems might lead to further improvements of the recognition accuracy.

7. Acknowledgements

This work was supported by Flemish Institute for the Promotion of Scientific and Technical Research in the Industry (contract SBO/40102).

8. References

[1] N. Chomsky and M. Halle, "The sound pattern of english," in *MIT Press*, 1968.

[2] A.V. Hansen, "Acoustic parameters optimised for recognition of phonetic features," in *Proc. Eurospeech*, 1997, pp. 397–400.

[3] K.-T. Lee and C. Wellekens, "Dynamic lexicon using phonetic features," in *Proc. Eurospeech*, Aalborg, Denmark, 2001, pp. 1413–1416.

[4] K. Elenius, "Phoneme recognition with an neural network," in *Proc. Eurospeech*, 1991, pp. 121–124.

[5] P. Dalsgaard, "Phoneme label alignment using acoustic-phonetic features and gaussian probability density functions," in *Computer, Speech and Language*, 1992, number 6, pp. 303–329.

[6] L. Deng and D.X. Sun, "A statistical approach to automatic speech recognition using the atomic speech units constructed from overlapping articulatory features," in *JASA*, May 1994, number 95 in 5, pp. 2702–2719.

[7] E. Eide, "Distinctive features for use in an automatic speech recognition system," in *Proc. Eurospeech*, Scandinavia, Aalborg, Denmark, 2001, pp. 1613–1616.

[8] S. Stüker, F. Metze, T. Schultz, and A. Waibel, "Integrating multilingual articulatory features into speech recognition," in *Proc. Eurospeech*, Geneva, 2003, pp. 1033–1036.

[9] F. Metze and A. Waibel, "A flexible stream architecture for asr using articulatory features," in *Proc. ICSLP*, Denver, Colorado, 2002.

[10] K. Kirchhoff, "Robust speech recognition using articulatory information," in *PhD Thesis*, Universität Bielefeld, 1999.

[11] F Stouten and J.-P. Martens, "On the use of phonological features for pronunciation scoring," in *Proc. ICASSP*, Toulouse, France, May 2006.

[12] S. King and P. Taylor, "Detection of phonological features in continuous speech using neural networks," in *Computer Speech and Language*, 2000, number 14 in 4, pp. 333–353.

[13] M.J.F. Gales, "Semi-tied covariance matrices for hidden markov models," in *IEEE Trans. on SAP*, May 1999, number 3 in 7, pp. 272–281.

[14] L. Josifovski, M. Cooke, P. Green, and A. Vizinho, "State based imputation of missing data for robust speech recognition and speech enhancement," in *Proc. Eurospeech*, 1999, vol. 6, pp. 2833–2836.

[15] S. Young, D. Kershaw, J. Odell, D. Ollasson, V. Valtchev, and P. Woodland, "The htk-book version 3.0," in *Cambridge University, Engineering Department*, 2000.