RELIABILITY-BASED ESTIMATION OF THE NUMBER OF NOISY FEATURES: APPLICATION TO MODEL-ORDER SELECTION IN THE UNION MODELS

Peter Jančovič, Münevver Köküer and Fionn Murtagh

School of Computer Science
Queen's University Belfast
Belfast BT7 1NN, Northern Ireland, UK
p.jancovic, m.kokuer, f.murtagh@qub.ac.uk

ABSTRACT

This paper is concerned with the multi-stream approach in speech recognition. In a given set of feature streams, there may be some features corrupted by noise. Ideally, these features should be excluded from recognition. To achieve this, a-priori knowledge about the identity, including both the number and location, of the noisy features is required. In this paper, we present a method for estimating the number of noisy feature streams. This method assumes no knowledge about the noise. It is based on calculation of the reliability of each feature stream and then evaluation of the joint maximal reliability. Since this method decreases the uncertainty about the noisy features and is statistical in nature, it can also be used to increase robustness of other classification systems. We present an application of this method to model-order selection in the union models. We performed tests on the TIDIGITS database, corrupted by noises affecting various numbers of feature streams. The experimental results show that this model achieves recognition performance similar to the one obtained with a-priori knowledge about the identity of the corrupted features.

1. INTRODUCTION

Speech signal may be represented by multiple feature streams, which may be obtained in general by using different sources of information or different processing techniques on a specific source. A specific case is the sub-band approach [1] [2], in which the full speech frequency-band is divided into several sub-bands, each sub-band being represented by an individual feature stream.

The multiple feature stream approach will usually tend to improve a speech recognition system, if the individual streams provide complementary information. Equally importantly, since in a given feature set some feature streams may be more robust than the others to a specific type of noise, the multi-stream approach can lead to a robust framework.

The key issue of the multi-stream approach is the formulation of the combination of feature streams. Ideally, those features that are unaffected or only slightly affected by noise should be selected, as they provide correct information about the utterance, whilst the features dominated by noise should be excluded as they can be detrimental to the recognition accuracy. This is the idea of missing feature theory. If a-priori knowledge about which of the features are affected by noise is available, this method may significantly improve the robustness of a speech recognition system [3]. However, in real world situations, this information is usually not available. Various methods were studied to identify the corrupted features, for example, in a multi-band recognition system, explicitly measuring the local signal-to-noise ratio (SNR), e.g. [4].

Recently, several studies have attempted to release the need for identification of the corrupted features. These include, for example, the full-combination model [5], the acoustic backing-off model [6], and the probabilistic union model [7]. The probabilistic union model combines the feature streams based on the probability theory for the union of random events. This approach assumes no knowledge about the location of the noisy feature streams. However, a parameter within this model that is called its *model-order* is related to the *number* of noisy feature streams, and its choice is critical to the recognition performance of the union model.

In this paper, we present a method for estimating the number of noisy feature streams in a given set of feature streams. This method assumes no knowledge about the noise. It is based on calculating the reliability of each feature stream and then evaluating the joint maximal reliability. We present an employment of the proposed reliability-based method to model-order selection in the union models. This model was tested using the TIDIGITS database, corrupted by frequency-localized noises affecting various numbers of feature streams. The results achieved indicate that the union model employed with the reliability-based model-order selection achieved recognition performance similar to the one obtained with full a-priori knowledge about the identity of the noisy feature streams.

2. RELIABILITY-BASED METHOD FOR ESTIMATING THE NUMBER OF NOISY FEATURE STREAMS IN A FEATURE SET

We consider that data is characterized by a set of N feature streams $o = (o_1, o_2, \dots , o_N)$; for instance, each individual feature stream o_n may characterize a different sub-band of the entire speech frequency band. In recognition, in a given a set of features, o, there may be some of the o_n's noisy, e.g. due to some unknown frequency-localized noise. We assume no knowledge about the noise. Firstly, we define the reliability of a feature stream. Then, the algorithm for estimating the number of noisy feature streams is presented.

2.1. Definition of reliability of a feature stream

Denote by $f(x_n|s)$ the probability distribution (or probability density function) associated with class s (e.g. an HMM state) that models feature stream o_n. For a given feature stream o_n, we denote by $r(o_n|s)$ a *reliability* of feature stream o_n at class s, which we define as

$$r(o_n|s) = \Big(\frac{f(o_n|s)}{\max_{x_n} f(x_n|s)}\Big)^{1/d} \qquad (1)$$

where d is dimension of feature streams o_n, and the exponent serves as a normalizing factor. The reliability $r(o_n|s)$ as defined in Eq. 1 expresses how different is the probability of a given feature stream o_n from the maximum possible probability. It produces, for each feature stream o_n, a reliability score within the interval $(0;1]$, allowing the scores to be consistent across different feature streams.

The value of a feature stream reliability close to 1 reflects a close similarity between the data and model; while reliability values approaching 0 mean little similarity between the data and model. As such, it is reasonable to assume that clean data on the correct model should produce reliability values which tend to be close to 1. On the other side, reliabilities of noisy feature streams, may become very small (i.e. approaching zero) on the correct model, because of the mismatch between the model and data; however, they may accidentally become high (i.e. close to one) on an incorrect model.

2.2. Algorithm description

Given a set of N feature streams, we calculate the reliability $r(o_n|s)$ of each feature stream o_n at each class s. This results in a set of reliabilities $\{r(o_1|s), \ldots, r(o_N|s)\}$ associated with each class s. Consider that the estimated number of corrupted feature streams, which we denote by a variable m, can be from 0 to $N-1$. For each m, we define a variable R_m, which we call an *average order-reliability*, as the geometric average of the maximal joint reliability of $(N-m)$ feature streams out of the entire set of N feature streams. Assuming independence between the feature streams, the R_m can be expressed as

$$\begin{aligned} R_m &= \Big[\max_{n_1,\cdots,n_{N-m}} r(o_{n_1} \wedge \ldots \wedge o_{n_{N-m}}|s)\Big]^{\frac{1}{N-m}} \\ &= \Big[\prod_{n=1}^{N-m} r(o_{(n)}|s)\Big]^{\frac{1}{N-m}} \qquad (2) \end{aligned}$$

where $r(o_{(n)}|s)$'s are the reliabilities, defined in Eq. 1, arranged in non-increasing order of magnitude, so that $r(o_{(1)}|s) \geq r(o_{(2)}|s) \geq \ldots \geq r(o_{(N)}|s)$.

Assume a situation when there are c feature streams corrupted by noise. Firstly, consider the case $m < c$. In this case, the average order-reliability R_m produced on the correct model will be much smaller than the average reliability obtained on the training (clean) data, because of including at least one noisy-feature stream reliability in the product operation. It is also reasonable to assume that as long as the test data do not resemble closely to an incorrect model, as a result of the product operation, the R_m produced on any incorrect model should also be smaller than the reliability obtained on the training data. For the case $m = c$, the average order-reliability R_m on the correct model eliminates c smallest reliabilities. Since the small feature-stream reliabilities on the correct model are considered to be the reliabilities of the noisy feature streams, it is reasonable to assume that in this situation R_m is the multiplication of only the reliabilities of $(N-c)$ uncorrupted feature streams. As such, it is reasonable to assume that this should produce a value that is similar to the one obtained on the training (clean) data.

Based on the above discussion, given a set of feature streams, the method for estimating the number of noisy (i.e. unreliable) feature streams, m^*, may be based on simple comparison of the value of each order-reliability R_m and some threshold γ. The threshold value corresponds to a reliability level below which we consider the data as unreliable and can be determined experimentally based on the training data. Algorithmic description of this method is depicted in Figure 1.

```
calculate $r(o_n|s)$  $\forall$ feature streams $o_n$,  $\forall$ classes $s$
for m=0 .. N-1
    $\forall s$, compute the order-reliability $R_m$
    if  $\exists s$ such that its associated $R_m > \gamma$
        $m^*$ = m;
        break;
    endif
endfor
```

Fig. 1. *An algorithmic description of the proposed method for estimating the number of unreliable feature streams.*

The following section discuss an application of this algorithm to the model-order selection in the union models.

3. APPLICATION TO MODEL-ORDER SELECTION IN THE UNION MODELS

We consider a conditional probability $P(o|s)$ of feature set $o = (o_1, o_2, \ldots, o_N)$ associated with class s. Assume that in the feature set o there are M feature streams corrupted by noise. Then we know that there exists one subset of $(N-M)$ features which are unaffected by noise. Combining these features by using the "and" (i.e. conjunction) operator derives a joint probability of the clean features, which should provide more discrimination than any of the marginal probabilities. Without knowing the location of the noisy features, the clean feature stream subset may be any of the subsets of $(N-M)$ feature streams. This uncertainty about the location of the noisy feature streams can be dealt with by using the "or" operator. As such, the useful information within the given feature set can be represented by combining the feature streams by the "and" and "or" operators. Then, the probability for the feature set o may be written as

$$P(o|s) = P\Big(\bigvee_{n_1,n_2,\cdots,n_{N-m}} o_{n_1}o_{n_2}\cdots o_{n_{N-M}}|s\Big) \qquad (3)$$

where $o_{n_1}o_{n_2}\cdots o_{n_{N-M}}$ is a subset in o containing $(N-M)$ feature streams which are combined with the "and" operator (for simplicity, the symbol $\wedge$ between the o_n's has been omitted), and the "or" operator $\vee$ is applied between all possible subsets of $(N-M)$ out of N feature streams, giving a total of ${}^{N}C_{N-M}$ combinations. Eq. 3 is called the *probabilistic union model of order M* [7], where the order of the model, M, takes a value in the range $0 \leq M \leq N-1$. As can be seen from the above discussion, to obtain optimal results, the model in Eq. 3 requires knowledge about the number of noisy feature streams. For example, in a simple case with three feature streams, Eq. 3 can take one of the following three possible forms, corresponding to order M=0, 1, and 2, respectively:

$$\begin{array}{llll}(M=0) & P(o|s) & = & P(o_1 o_2 o_3|s) \\ (M=1) & P(o|s) & = & P(o_1 o_2 \vee o_1 o_3 \vee o_2 o_3|s) \\ (M=2) & P(o|s) & = & P(o_1 \vee o_2 \vee o_3|s)\end{array}$$

The form (M=0) is best suited to the situations in which all the feature streams are reliable (i.e. no corruption). Forms (M=1) and (M=2) are best suited to the situations in which there is one and two noisy feature streams, respectively. For example, in form (M=1) assuming one noisy feature stream, the union of the three conjunctions will include one conjunction providing the joint probability of the remaining two clean feature streams; the other two conjunctions each contain a noisy feature stream, with a correspondingly low probability on the correct model, and therefore make only a small contribution to the union probability associated with the correct model.

Without any knowledge about the noise, we face the problem of selecting the model-order of the union model, in order to obtain optimal recognition performance. In [8], an algorithm for model-order selection in the union models based on state-duration pattern has been proposed. Since this method is based on a duration principle, the model-order can only be selected on an utterance level. Here, we employed the reliability-based method described in Section 2 for selecting the model-order. At each frame time, given a set of features, the reliability method is applied to estimate the number of noisy features; this determines the order of the union model that is used for combining the features. The advantage of using the proposed reliability method over the state-duration method is that the model-order can be selected on a frame level. As such, when using the reliability method we should be capable of dealing more effectively with noises that cause that the number of noisy feature streams varies over time.

4. EXPERIMENTS AND RESULTS

Experiments have been carried out using the isolated-digit part of the TIDIGITS database. This database includes eleven isolated-digit words: “one” to “nine”, “zero”, and “oh”, each digit surrounded by silence parts.

The speech signal, sampled at 8 kHz, is divided into frames of 30 ms, with an overlap of 10 ms between frames. Both pre-emphasis and Hamming window are applied to each frame. For each frame, a multi-channel, Mel-scaled filter bank analysis with 35 channels is used to estimate the log-amplitude spectra of the speech. These filter channels are then grouped uniformly into 5 sub-bands, each sub-band consisting of information from 7 channels. A DCT is applied within each sub-band and the first 4 MFCCs coefficients form the sub-band feature vector. In order to include dynamic spectral information, the first-order delta parameters were calculated and added to each sub-band feature vector. The probabilities of these five individual feature streams are merged at the frame level using the probabilistic union model equipped with the reliability-based method for estimating the model-order. A 12-state HMM is estimated for each word, with the first and last states being tied among all the vocabulary words to account for the silence parts of the utterances. The training of HMMs was performed on clean utterances from the training set. For recognition, the testing set was corrupted by various types of noises, which consisted of frequency-localized noise component(s). The noise was added to the speech signal. The frequency-localized noise was generated by passing the Gaussian white noise through a band-pass filter. The 3dB cut-off bandwidth of the noise was fixed at 100 Hz and the central frequency of the noise varied. In particular, five different central frequencies were chosen, which are 600 Hz, 1000 Hz, 1500 Hz, 2100 Hz and 2800 Hz. The calculation of the SNR was based on the averaged energy of all the test speech utterances; so the noise in each utterance is of a constant loudness, regardless of the actual energy of speech in that utterance. Two SNR conditions were considered, i.e. SNR=10dB and SNR=0dB.

4.1. Determining the threshold γ

Firstly, in order to determine the threshold value γ, we performed experiments on the training set corrupted by various frequency-localized noises. We tested different threshold values within a range $(0.45, 0.75)$. Figure 2 shows the recognition results achieved. As can be seen from Figure 2, the recognition performance shows similar behaviour for different SNR and different number of feature streams being corrupted. Based on these results, we set the threshold $\gamma = 0.63$ for all the experiments presented below.

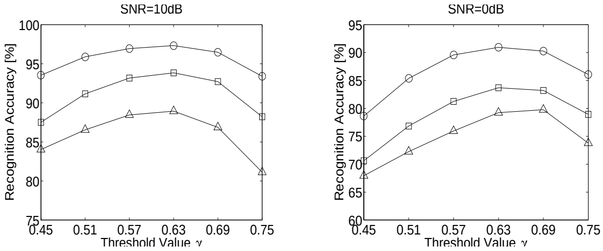

Fig. 2. *Recognition accuracy results on the training set corrupted by various frequency-localized noises as a function of threshold value γ. Number of noisy features depicted as: o - 1, □ - 2, △ - 3.*

4.2. Experiments with the numbers of corrupted features fixed over time

In this section, we present experiments with noises that corrupt the same number of feature streams over the entire utterance.

The frequency-localized noise components mentioned above were chosen to create an effect that there were one sub-band, two sub-bands, and three sub-bands corrupted. Specifically, the noises with central frequencies 600 Hz, 1000 Hz, 2100 Hz and 2800 Hz were located within sub-band 2, 3, 4 and 5, respectively, and each thus caused only corruption of one sub-band (i.e. one feature stream). The noise with central frequency 1500 Hz was located around the border of sub-bands 3 and 4, and thus causing the corruption of two feature streams. The noises corrupting two feature streams were also created by combination of two noise components with different central frequencies, in particular, 600 Hz and 1000 Hz, 600 Hz and 2100 Hz, 600 Hz and 2800 Hz, 1000 Hz and 1500 Hz, 1000 Hz and 2100 Hz, 1000 Hz and 2800 Hz, 1500 Hz and 2100 Hz, 2100 Hz and 2800 Hz. The noises consisting of components 600 Hz and 1500 Hz, 1500 Hz and 2800 Hz caused corruption of three feature streams.

Table 1 presents the recognition results obtained by the union model with all the model-orders (i.e. $M = 0, \ldots, 4$) within our five-band system. The results are shown for both the clean and noisy speech, as a function of the SNR and the number of corrupted feature streams (nC). From Table 1 it can be seen that the

Table 1. *Recognition accuracy results in the recognition system with five feature streams by the union model with all model-orders and with the order selected by using the reliability-based method.*

SNR		Union model					
(dB)	nC	with order M					rel.-based
		0	1	2	3	4	order sel.
clean	0	**99.2**	99.1	98.4	96.1	83.5	99.1
10	1	88.3	**97.4**	96.5	91.3	75.6	96.7
	2	75.6	89.6	**93.6**	88.3	72.0	92.9
	3	72.3	79.0	83.1	**83.9**	69.1	86.7
0	1	67.7	**91.7**	89.4	82.2	67.2	90.0
	2	50.7	69.9	**83.0**	77.4	62.2	81.9
	3	52.2	63.0	70.1	**77.8**	62.8	76.3

union model obtained optimal recognition performance (shown in bold) when the model-order equals the number of noisy feature streams. On the right side in Table 1 are shown the recognition results obtained when the reliability-based algorithm was employed for automatic order selection. We can see that these recognition results are similar to the results obtained by assuming that the number of noisy features is known a-priori.

4.3. Experiments with the numbers of corrupted features varied over time

Next, we performed experiments when the number of corrupted feature streams varies throughout the utterance. Specifically, three different noises with frequency-characteristics depicted in Figure 3 were used.

The recognition results are presented in Table 2. The first model we compared was a missing-feature model, which assumed full a-priori knowledge of the corrupted feature streams (i.e. the number and location of corrupted feature streams) and removed those features manually from recognition. The second model being compared was the baseline HMM, which combines the feature streams only by the "and" operator. The third model being compared was the union model with state-duration method for model-order selection [8]. As can be seen from Table 2 the union model equipped with the reliability-based order selection method significantly outperformed over both the baseline HMM, and also the union model equipped with the state-duration method for selecting the order. As discussed earlier, this is because the reliability-based method can estimate the order on each frame basis, which is not possible when using the state-duration method.

Table 2. *Recognition accuracy results in the recognition system with five feature streams by the union model with order selection based on the state-duration method and reliability method.*

SNR	Noise	Model			
(dB)	type	A-priori	Baseline	Union with order alg.	
		knowled.	HMM	state-dur.	rel.-based
10	N1	95.0	76.3	88.9	93.0
	N2	95.3	72.4	90.8	93.1
	N3	95.2	72.8	90.1	93.6
0	N1	87.6	53.1	77.3	80.9
	N2	87.5	50.5	83.2	88.5
	N3	90.4	45.1	82.8	90.8

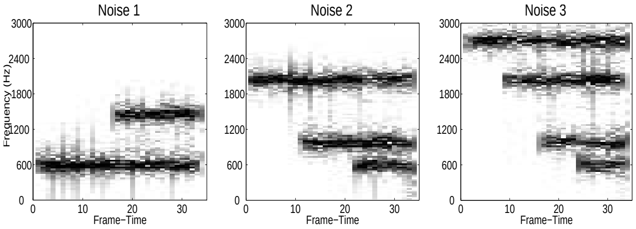

Fig. 3. *The time-frequency characteristics of the non-stationary noises used.*

5. CONCLUSION

In this paper, we presented a reliability-based method for estimating the number of corrupted feature streams in a given set of feature streams. This method assumes no knowledge about the noise. We presented an application of this method to model-order selection in the union models. The experiments were performed on the TIDIGITS database corrupted by noises affecting various numbers of feature streams. Significantly improved results in comparison to the baseline HMM and the previous union model equipped with state-duration method for order selection were obtained. Indeed, in many cases, the proposed model obtained recognition performance similar to model with full a-priori knowledge about the noisy feature streams. The proposed reliability-based method for estimating the number of noisy feature streams is general and thus can be used to improve the robustness of other classification systems.

This work was supported by UK EPSRC grant GR/M93734.

6. REFERENCES

[1] S. Tibrewala and H. Hermansky, "Sub-band based recognition of noisy speech," *ICASSP, Munich*, pp. 1255–1258, 1997.

[2] H. Bourlard and S. Dupont, "A new ASR approach based on independent processing and recombination of partial frequency bands," *ICSLP, Philadelphia, USA*, 1996.

[3] R.P. Lippmann and B.A. Carlson, "Using missing feature theory to actively select features for robust speech recognition with interruptions, filtering and noise," *Eurospeech, Rhodes, Greece*, pp. 37–40, 1997.

[4] A. Drygajlo and M. El-Maliki, "Speaker verification in noisy environment with combined spectral subtraction and missing data theory," *ICASSP, Seattle, WA*, vol. I, pp. 121–124, 1998.

[5] A. Morris, A. Hagen, and H. Bourlard, "The full combintion sub-bands approach to noise robust HMM/ANN based ASR," *Eurospeech, Budapest, Hungary*, pp. 599–602, 1999.

[6] J. de Veth, B. Cranen, and L. Boves, "Acoustic backing-off in the local distance computation for robust automatic speech recognition," *ICSLP, Sydney, Australia*, pp. 65–68, 1998.

[7] J. Ming and F.J. Smith, "A probabilistic union model for sub-band based robust speech recognition," *ICASSP, Istanbul, Turkey*, pp. 1787–1790, 2000.

[8] P. Jančovič and J. Ming, "A probabilistic union model with automatic order selection for noisy speech recognition," *Journal of the Acoustical Society of America*, vol. 110, pp. 1641–1648, 2001.