APPLICATION OF THE DEMPSTER-SHAFER THEORY OF EVIDENCE TO IMPROVED-ACCURACY ISOLATED-WORD RECOGNITION

R.L. Bowles and R.I. Damper

Department of Electronics and Computer Science, University of Southampton
Highfield, Southampton SO9 5NH, U.K.

Abstract

This paper describes experiments in which the outputs of three isolated-word recognition algorithms were combined to yield a lower average error rate than that achieved by any individual algorithm. The input tokens were simulated, fixed-length spectral speech patterns subjected to additive noise and either a "high" or "low" degree of time distortion. The three recognition techniques used were dynamic time warping, hidden Markov modelling and a multi-layer perceptron. Two combination techniques were employed: the formula derived from the Dempster-Shafer (D-S) theory of evidence and simple majority voting (MV). D-S performed significantly better than MV under both time-distortion conditions. Evidence is also presented that the assumption of independent word scores, which is necessary for D-S theory to be strictly applicable, is questionable.

1 INTRODUCTION

The problem domains of interest in artificial intelligence are characterised by the need to handle uncertain information. Several standard methodologies for this have arisen, each with its own representation of the degree of uncertainty of the data, including Bayesian updating, belief functions [1], fuzzy logic [2] and the MYCIN calculus [3]. Each approach gives a way of combining evidence from distinct knowledge sources; a comprehensive survey of the methods is presented in [4,5]. Apart from our own work [6], these techniques have not (as far as we are aware) been employed in automatic speech recognition.

According to Allerhand [7], the use of simple models in speech recognition creates an inherent performance limitation. A plateau is reached, set by the assumptions implicit in the model, where further improvement cannot be made. Combination of evidence offers a possible means of overcoming the fundamental limitation imposed by use of a single model. In [6], we showed that it was possible in certain circumstances to obtain a significant increase in recognition accuracy by combining the outputs from two or three distinct isolated-word recognition (IWR) algorithms subjected to the same simulated speech input using Dempster-Shafer (D-S) theory (see below). In this case, the individual algorithms act as separate sources of evidence (assumed independent) concerning word identity. The algorithms used at that time were: dynamic time warping (DTW), hidden Markov modelling (HMM) and a rather simple spectral peak-picking technique (SPP). No real efforts were made to optimise the separate algorithms; indeed, our position was that the combination approach should, if anything, compensate for their imperfections.

In this paper, we report results obtained when three independent algorithms all achieve a high degree of accuracy (typically in the range 90 to 99%). To effect this, we have improved considerably the training of the HMM word models and used the more powerful multi-layer perceptron (MLP) in place of SPP. We have also greatly improved the method of obtaining belief functions from distance scores and this is described. We also compare results using D-S combination with those obtained using the simplest possible combination technique, namely majority voting (MV).

2 DEMPSTER-SHAFER COMBINATION

The D-S formula [8] is a means of combining evidence based on belief functions so as to select among competing hypotheses. In this theory, the frame of discernment, \( \Theta \), is a set of exhaustive and mutually exclusive hypotheses (singletons). The set of possible hypotheses is the powerset of \( \Theta \), including the empty set, called \( \Omega \). The subsets \( A \) of \( \Omega \) for which there is...
direct evidence are termed the focal elements of $\Omega$.

### 2.1 Probability masses and belief

Evidence for the hypotheses takes the form of probability assignments — called basic probability masses, $m(A)$ — to each of the focal elements, $A$. The masses are probability-like in that they are in the range $[0,1]$ and sum to 1 over all hypotheses, but are not exactly probabilities; rather, they represent the belief assigned to a focal element, however measured. The belief in an hypothesis $H$ is termed $BEL(H)$ or, sometimes, the lower probability of $H$. It is equal to the sum of all the probability masses of the subsets of $H$:

$$BEL(H) = \sum_{A \subseteq H} m(A) \quad (1)$$

The plausibility of $H$, sometimes called the upper probability, is defined to be $(1 - BEL(H))$. It can be considered as the extent to which the evidence does not contradict the hypothesis. Belief values always lie in the range 0 to 1; $BEL(H) = 1$ means that $H$ is effectively certain while $BEL(H) = 0$ means that there is a total lack of belief in $H$ (not to be confused with disbelief).

When the hypotheses, $\Omega$, are all singletons, belief and plausibility become identical.

The combination formula allows for two different sets of probability masses from independent sources, but relevant to the same hypotheses, to be combined to give overall belief values. If $F$ is an hypothesis and $m(G)$ and $m(H)$ are the probability assignments to focal elements $G$ and $H$, then the combined evidence for $F$ is given by:

$$BEL(F) = \frac{\sum_{G \cap H = F} m(G) \cdot m(H)}{\sum_{\emptyset \neq G} m(G) \cdot m(H)} \quad (2)$$

The denominator acts as a normalising term to ensure that the combined belief value lies in the range 0 to 1. The combination is commutative and associative and, hence, any number of masses can be combined in any order.

### 2.2 Isolated word recognition

In the case of IWR, the hypotheses, $\Omega$, are the possible identities of the words and the evidence for these words is the similarity scores corresponding to them. The hypotheses reduce to a set of singletons since each score obtained relates only to a single word. That is, for a vocabulary of $N$ words, there are $N$ hypotheses: $H_i$ is “the $i$th word of the vocabulary was spoken", where $1 \leq i \leq N$. In this case, $\Omega$ is actually identical to the frame of discernment, $\Theta$.

(Note that we have ignored the possibility of out-of-vocabulary utterances, corresponding to the inclusion of the empty set in $\Omega$.) Thus, the only set intersections which are non-empty are those pertaining to the probability masses for the $i$th word obtained from the three algorithms, $X$, $Y$ and $Z$, say. Thus, (2) becomes:

$$BEL(H_i) = \frac{m(x_i) \cdot m(y_i) \cdot m(z_i)}{\sum_{j=1}^{N} m(x_j) \cdot m(y_j) \cdot m(z_j)} \quad (3)$$

Applying the D-S formula in the form of (3), i.e. with singletons, is very close to Bayesian updating but with important differences. First, probability masses and a belief appear in place of probabilities and, second, an assumption is made about independence of the evidence (scores) which is not essential to the Bayesian calculus [5]. Because of the way we derive probability masses from scores (see below), which means they are not true probabilities, we feel it is preferable to view our method in a D-S framework.

### 2.3 Probability mass computation

A method of converting the similarity scores into probability masses is required for use in (3). According to Lindley [9], probability is the best possible measure of belief in an hypothesis. Further, if probabilities were available, we could use Bayes’ rule to determine the “belief” in hypothesis, $H$, given the evidence available, $E$:

$$p(H|E) = \frac{p(E|H) \cdot p(H)}{p(E)} \quad (4)$$

Since all “words” are equally likely in our simulation, and there are no out-of-vocabulary utterances, $p(H)$ is simply equal to $1/N$. However, the probabilities $p(E|H)$ and $p(E)$ are effectively unknown; but, we can estimate them from the distribution of scores as we now describe.

2,000 tokens were input to the three recognition algorithms and scores calculated for each token matched against all the word models. For each algorithm, a distribution was constructed of all scores obtained; these were then assumed to be Gaussian. Thus, for any particular score, $E$, an estimate of $p(E)$ can be obtained using the normal formula. In like manner, $p(E|H)$ can be estimated for each word by constructing a distribution of correct match scores only; there are 32 of these distributions for each algorithm. This then allows us to estimate $p(H|E)$ using (4). Since the result is only probability-like, we prefer to think of it as a probability mass, $m(H)$. This
method of computing probability masses ensures that all the \( m(H) \)'s sum to 1.

3 THE TEST SYSTEM

The test system consisted of the three standard recognition algorithms – DTW, HMM and MLP – each of which was subjected to simulated speech patterns drawn from a “vocabulary” of \( N = 32 \) words. The outputs from these algorithms were then combined to produce an overall score for each word.

3.1 Simulated speech data

Because of the large amount of input data needed to test the combination strategy, we have chosen to use simulated speech. Each token is intended to mimic the output from a 16-filter bandpass analysis, time-normalised to a sequence of 16 frames. 32 “prototypes” were manually created to resemble actual patterns. Further tokens were produced by randomly adding and deleting frames and re-normalising; Gaussian noise was also added to the spectral values to produce a 16 dB SNR. Two degrees of time distortion were applied: low (LTD) in which each frame had a 33% chance of being repeated or deleted, and high (HTD) in which the chance was 67%.

3.2 Recognition algorithms

The DTW algorithm was exactly as described in [6]. This used a conventional asymmetric local path constraint based on the Chebychev distance metric. Warping paths were constrained to end at the final frame in both the test word and the vocabulary prototype. Global paths were constrained by setting a semi-window width of 5.

As in [6], the HMMs used were discrete, 5-state, left-to-right, double-step models. However, the models used here were trained on synthetic tokens. (produced as above), the frames of which were clustered into 32 classes using the c-means clustering algorithm in conjunction with a Euclidean distance measure. The figure of 32 cluster centres was chosen empirically to maximise the HMM recognition rate. Training employed the Baum-Welch algorithm; the forward-backward procedure was used for matching, producing a probability reflecting the degree of similarity.

For the MLP, a 3-layer network was used with 256 nodes in the input layer, 20 nodes in the hidden layer and 32 in the output layer. The function of the input layer was to hard-limit the speech-pattern values to 0 or 1. The nodes in the hidden layer were not fully connected to the input layer but, rather like the “zonal units” of [10], were divided into 4 groups of 5. Each group of 4 was connected to 4 contiguous frequency bands in the input. This was intended primarily to reduce computation time but it also slightly mimicked the critical band filtering of the auditory system [10]. The output layer was fully interconnected with the hidden layer and each output node corresponded to one vocabulary word. The MLP was trained using error-back propagation by clamping one output high when the corresponding word pattern was presented to the inputs. In practice, a test word caused all the outputs to go to different values between 0 and 1, these values being taken as the vector of similarity scores for the MLP.

3.3 Combination techniques

The D-S formula, equation (3), was used to obtain \( BEL \) values from the probability masses found as in Section 2.3. The word recognised was that with the highest \( BEL \) value.

For reference purposes, we also combined the outputs from the three individual algorithms using majority voting with each algorithm contributing a single vote on word identity. This is perhaps the simplest combination strategy possible. Further, it is amenable to theoretical analysis under an independence assumption so that predicted and obtained accuracies can be compared. The degree of differences between these two can be taken as an indication of the validity of the assumption.

Taking \( P, Q \) and \( R \) to be the recognition rates of the three algorithms, and further taking these to be synonymous with the probability of correct recognition, the combined accuracy with majority voting, \( MV \), is easily shown to be:

\[
MV = PQ + PR + QR - 2PQR
\]

Error rates were defined as \( 1 - MV \), i.e they include both false acceptances and rejections (which occur when all three algorithms vote differently.)

4 RESULTS

2,000 tokens of simulated speech data were input to the three recognition algorithms, \( BEL \) values computed and the rule of combination applied to produce a composite score, the maximum value of which indicated the word recognised from the 32 candidates. This was repeated 10 times for both high- and low-time distortion in order to give some idea of the vari-
This work has shown that Dempster-Shafer combination can be used to increase the accuracy of isolated word recognition over that obtainable from a single algorithm even when the individual algorithms each achieve high accuracy.

Although the speech data in these trial runs was synthetic, efforts were made to ensure the test words were realistic. Our next step is to see if a similar improvement is possible using real speech, at least for the prototypes. We are also currently implementing something closer to Bayesian updating which avoids the independence assumption.

In the long term, we foresee combination techniques as contributing to connected-word and large-vocabulary recognition based on sub-word modelling rather than whole-word pattern matching. In this case, the individual sources will provide evidence for the identity of sub-word units.

6 REFERENCES