IMPROVING UTTERANCE VERIFICATION USING A SMOOTHED NAIVE BAYES MODEL

Alberto Sanchis, Alfons Juan and Enrique Vidal

Institut Tecnològic d'Informàtica
Departament de Sistemes Informàtics i Computació
Universitat Politècnica de València
46071, València (Spain)
{asanchis,ajuan,evidal}@iti.upv.es

ABSTRACT

Utterance verification can be seen as a conventional pattern classification problem in which a feature vector is obtained for each hypothesized word in order to classify it as either correct or incorrect. It is unclear, however, which predictor (pattern) features and classification model should be used. Regarding the features, we have recently proposed a new feature, called *Word Trellis Stability* (WTS), that can be profitably used in conjunction with more or less standard features such as *Acoustic Stability*. This is confirmed in this paper, where a *smoothed naive Bayes* classification model is proposed to adequately combine predictor features. On a series of experiments with this classification model and several features, we have found that the results provided by each feature alone are outperformed by certain combinations. In particular, the combination of the two abovementioned features has been consistently found to give the most accurate result in two verification tasks.

1. INTRODUCTION

Current speech recognition systems are not error-free and, in consequence, it is desirable for many applications to predict the reliability of each hypothesized word. From our point of view, this can be seen as a conventional pattern recognition problem in which each hypothesized word is to be transformed into a feature vector and then classified as either correct or incorrect [1]. The basic problem then is to decide which predictor (pattern) features and classification model should be used.

We have recently proposed a new feature, called *Word Trellis Stability* (WTS) [2], that performs relatively well in comparison with several well-known features [3, 4, 5, 6]. In this paper, we propose a *smoothed naive Bayes* classification model to profitably combine these features. The model itself is a combination of *word-dependent* (specific) and *word-independent* (generalized) naive Bayes models. As in statistical language modelling, the purpose of the generalized model is to smooth the (class posteriors) estimates given by the specific models.

The smoothed naive Bayes model and a brief review of predictor features are given in sections 2 and 3, respectively. In section 4, experimental results are reported on two verification tasks.

2. SMOOTHED NAIVE BAYES MODEL

We denote the class variable by c; $c = 0$ for correct and $c = 1$ for incorrect. Given a hypothesized word w and a D-dimensional vector of (discrete) features $\boldsymbol{x}$, the class posteriors can be calculated via the Bayes' rule as

$$P(c|\boldsymbol{x}, w) = \frac{P(c|w)\, P(\boldsymbol{x}|c, w)}{\sum_{c'} P(c'|w)\, P(\boldsymbol{x}|c', w)} \qquad (1)$$

Therefore, our basic problem is to estimate $P(c|w)$ for each word and $P(\boldsymbol{x}|c, w)$ for each class-word pair. For simplicity, we make the naive Bayes assumption that the features are mutually independent given a class-word pair,

$$P(\boldsymbol{x}|c, w) = \prod_{d=1}^{D} P(x_d|c, w) \qquad (2)$$

Given N training samples $\{(\boldsymbol{x}_n, c_n, w_n)\}_{n=1}^{N}$, we can estimate the unknown probabilities using the conventional frequencies

$$P(c|w) = \frac{N(c, w)}{N(w)} \qquad (3)$$

$$P(x_d|c, w) = \frac{N(x_d, c, w)}{N(c, w)} \qquad d = 1, \ldots, D \qquad (4)$$

where the $N(\cdot)$ are suitably defined event counts; i.e., the events are (c, w) pairs in (3) and (x_d, c, w) triplets in (4).

Unfortunately, these frequencies often underestimate the true probabilities involving rare words and the incorrect class.

THIS WORK WAS PARTIALLY SUPPORTED BY THE EU PROJECT "TT2" (IST-2001-32091).

To circumvent this problem, we have considered an *absolute discounting* smoothing model imported from statistical language modelling [7]. The idea is to discount a small constant $b \in (0,1)$ to every positive count and then distribute the gained probability mass among the null counts (unseen events). Thus, for each word w, if $N(c,w) = 0$ for $c = 1$ (or $c = 0$), (3) is replaced by

$$P(c|w) = \begin{cases} \dfrac{N(c,w) - b}{N(w)} & \text{if } N(c,w) > 0 \\ \dfrac{b}{N(w)} & \text{if } N(c,w) = 0 \end{cases} \quad (5)$$

Similarly, for each (c,w), if $N(x_d,c,w) = 0$ for one or more possible values of x_d, the probability function (4) becomes

$$P(x_d|c,w) = \begin{cases} \dfrac{N(x_d,c,w) - b}{N(c,w)} & \text{if } N(x_d,c,w) > 0 \\ M \dfrac{P(x_d|c)}{\sum\limits_{x'_d : N(x'_d,c,w)=0} P(x'_d|c)} & \text{if } N(x_d,c,w) = 0 \end{cases} \quad (6)$$

where M denotes the gained probability mass ($\frac{b}{N(c,w)}$ times the number of seen events). Note that $P(x_d|c)$ is used as a *generalized distribution* to divide M among the unseen events. To prevent null estimates, it is also smoothed by absolute discounting (with a uniform backoff)

$$P(x_d|c) = \begin{cases} \dfrac{N(x_d,c) - b}{N(c)} & \text{if } N(x_d,c) > 0 \\ \dfrac{b}{N(c)} \dfrac{\sum 1}{\sum\limits_{x'_d : N(x'_d,c)=0} 1} & \text{if } N(x_d,c) = 0 \end{cases} \quad (7)$$

In practice, there are many (c,w) pairs for which nearly all $N(x_d,c,w)$ counts are null and, therefore, even the smoothed model (6) gives inaccurate estimates. To deal with these extreme cases, we have defined a global threshold for the $N(c,w)$ counts. For those (c,w) pairs with counts below this threshold, the generalized model (7) is used instead of (6). Similarly, if a word w does not occur in the training data, $P(c|w)$ is approximated by $P(c)$.

Using the models trained as explained above, in the test phase, utterance verification is performed by classifying a word as incorrect if $P(c = 1 \mid \boldsymbol{x}, w)$ is greater that a certain threshold τ (cf. section 4.3).

3. PREDICTOR FEATURES

A set of well-known features has been selected to perform the experiments presented in section 4:

- *Acoustic stability*: Number (or percentage) of times that a hypothesized word appears at the same position (as computed by Levenshtein alignment) in K alternative outputs of the speech recognizer obtained using different values of the *Grammar Scale Factor* (GSF), i.e. a weighting between acoustic and language model scores [3].
- *LMProb*: Language model probability [4].
- *Hypothesis density (HD)*: The average number of the active hypotheses within the hypothesized word boundaries [5].
- *PercPh*: The percentage of hypothesized word phones that match the phones obtained in a "phone-only" decoding [4].
- *Duration*: The word duration in frames divided by its number of phones [4].
- *ACscore*: The acoustic log-score of the word divided by its number of phones [6].

In addition we consider a new feature that we have recently introduced called "Word Trellis Stability" (WTS). Let w be a word of the recognized sentence and let s_w, e_w be the starting and ending frames of w, $0 \leq s_w < e_w < N$, where N is the number of frames of the given utterance. The WTS of w is computed as:

$$WTS(w) = \frac{1}{e_w - s_w + 1} \sum_{t'=s_w}^{e_w} \frac{C(w,t')}{\sum\limits_{w'} C(w',t')}$$

$$C(w,t') = \sum_{t=t'}^{N-1} \sum_{h \in \mathcal{H}_t(w,t')} (\alpha_f - \alpha_i)$$

where $\mathcal{H}_t$ is a set of word-boundary partial hypotheses that are most probable at time *t* for a certain range of GSF values [α_i,α_f]. In addition, in each hypothesis of $\mathcal{H}_t(w,t')$ the word w must be active at time frame t'. More details about the WTS can be found in [2].

4. EXPERIMENTS

We carried out experiments using two different corpora. One is the *Traveler task*, a Spanish speech corpus of person-to-person communication utterances at the reception desk of a hotel [8]. The other is the *FUB task*, an Italian speech corpus of *phone* calls to the front desk of a hotel [9]. Main features of the (disjoint) training and test sets, for both corpora, acquired in the context of the EuTrans project [8, 9], are summarized in table 1.

Table 1. Traveler and FUB speech corpus

	Traveler task*		FUB task	
	training	test	training	test
# speakers	20	12	276	24
# run. words	13,728	3,390	52,511	5,381
# vocabulary	683	–	2,459	–
bigram perplex.	–	6.8	–	31

(*) The training/test partition is slightly different from the one used in [8]: here the test-set speakers are a subset of the training-set speakers, but the utterances differ.

4.1. Traveler task

For the experiments with the *Traveler task* speech corpus, 24 context-independent Spanish phonemes were modeled by conventional left-to-right continuous-density hidden Markov models (HMM). A bigram language model was estimated using the whole training *text* corpus of the *Traveler task* [8]. The test-set Word Error Rate was 5.5 %.

4.2. FUB task

The *FUB* corpus involves highly spontaneous speech data and contains many non-speech artifacts. The training set was used to train Italian context-dependent phone models. The acoustic models were left-to-right continuous density HMMs, trained using Linear discriminant analysis (LDA) and a Viterbi approximation [10]. Decision-tree clustered generalized triphones (CART with 1,500 tied states plus silence) were used as phone-units. A smoothed trigram language model was estimated using the transcription of the training utterances. The test-set Word Error Rate was 27.5 %.

4.3. Experimental results

To perform the experimental study, a conventional continuous speech recognizer based on Viterbi beam search has been used with the language and acoustic models described in the last subsections.

In evaluating verification systems, two measures are of interest: the *True Rejection Rate* (TRR, the number of words that are incorrect and are classified as incorrect divided by the number of words that are incorrect) and the *False Rejection Rate* (FRR, the number of words that are correct and are classified as incorrect divided by the number of words that are correct). The trade-off between TRR and FRR values depends on a decision threshold τ (see section 2). A *Receiver Operating Characteristic* (ROC) curve represents TRR against FRR for different values of τ. The area under a ROC curve divided by the area of a worst-case diagonal ROC curve, provides an adequate overall estimation of the classification accuracy. We denote this area ratio as AROC. Note that an AROC value of 2.0 would indicate that all words can be correctly classified. We have used both ROC curves and the AROC measure to conveniently evaluate and compare the classification accuracy for different feature combinations.

Table 2 shows the AROC value using the (single-feature) smoothed model (eq. 1) for the *Traveler* and the *FUB* corpus. It can be observed that *acoustic stability* (AS) and WTS are consistently the best performing single features for both corpus. On the other hand, the newly introduced WTS feature significantly outperforms all the other traditional features.

Table 2. AROC value for each individual feature

Traveler task		*FUB task*	
Feature	AROC	Feature	AROC
AS	1.73	AS	1.69
WTS	1.73	WTS	1.62
HD	1.65	LMProb	1.55
PercPh	1.65	HD	1.49
LMProb	1.62	PercPh	1.44
Duration	1.62	ACscore	1.42
ACscore	1.59	Duration	1.41

To further exploit the usefulness of the features, the smoothed naive Bayes model presented in section 2 was used to combine different features in the classification process. Classification accuracy improvements were obtained through different feature combinations. In the *Traveler* corpus, a some improvement is achieved when the two best single features are combined. None of the other combinations outperforms the AS and the WTS single-feature performance. For the *FUB* corpus, different combinations of the three best single-features work better than the single-feature performance. A significant improvement is achieved when WTS is combined with AS. The combination of LMProb and AS performs slightly better than the AS-only performance. No three-feature combination improves the combination of AS and WTS. Table 3 shows the AROC values for the best results obtained.

Table 3. AROC values for the best feature combinations

Traveler task	
Feature	AROC
AS+WTS	1.77

FUB task	
Feature	AROC
AS+WTS	1.75
AS+LMProb	1.71

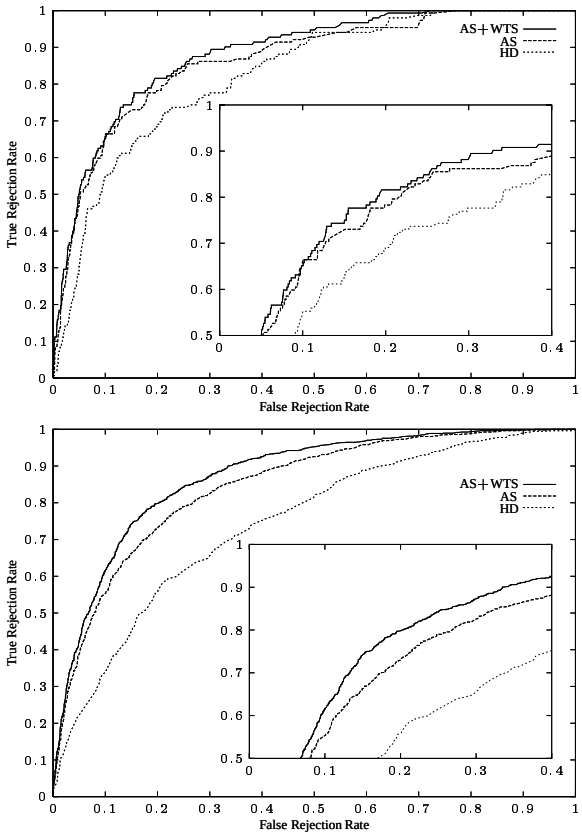

Fig. 1. Comparative ROC curves for single features AS and HD versus the best feature combination (AS+WTS). Top: *Traveler* corpus; bottom: *FUB* corpus.

As a summary of the results, figure 1 shows ROC curves obtained using the best feature combination, the best single feature and a traditional feature (HD), for the *Traveler* and the *FUB* corpus.

Our best results are based on two main contributions. The first one is the discrimination power of the newly introduced WTS feature. And the second is the capability of the proposed naive Bayes model to improve the performance of individual features, by adequately combining them under a sound statistical framework.

5. CONCLUSIONS

We have proposed a smoothed naive Bayes model to estimate confidence measures in speech recognition verification. Smoothing is based on traditional techniques applied in the context of statistical language modelling for speech recognition. The results show that the combination of different features is significantly better than the single-feature performance of a set of well-known features. Also, the WTS feature, which we have recently introduced [2], has demonstrated to be particularly effective to improve the classification performance.

6. REFERENCES

[1] A. Sanchis, V. Jiménez, and E. Vidal, "Efficient Use of the Grammar Scale Factor to Classify Incorrect Words in Speech Recognition Verification," in *ICPR*, 2000, vol. 3, pp. 278–281.

[2] A. Sanchis, A. Juan, and E. Vidal, "Estimating confidence measures for speech recognition verification using a smoothed naive bayes model," Submitted to IbPRIA'2003.

[3] T. Zepenfeld, M. Finke, K. Ries, M. Westphal, and A. Waibel, "Recognition of conversational telephone speech using the JANUS speech engine," in *ICASSP*, 1997, pp. 1815–1818.

[4] L. Chase, *Error-responsive feedback mechanisms for speech recognizers*, Ph.D. thesis, School of Computer Science, Carnegie Mellon University, USA, 1997.

[5] T. Kemp and T. Schaaf, "Estimating confidence using word lattices," in *EUROSPEECH*, 1997, pp. 827–830.

[6] T. Schaaf and T. Kemp, "Confidence measures for spontaneous speech recognition," in *ICASSP*, 1997, pp. 875–878.

[7] H. Ney, S. Martin, and F. Wessel, "Statistical language modeling using leaving-one-out," *Young, S. and Bloothoft, G., editors, Corpus Based Methods in Language and Speech Processing*, pp. 174–207, 1997.

[8] J.C. Amengual, J.M. Benedí, F. Casacuberta, M.A. Castaño, A. Castellanos, V.M. Jiménez, D. Llorens, A. Marzal, M. Pastor, F. Prat, E. Vidal, and J.M. Vilar, "The EuTrans-I speech translation system," *Machine Translation*, vol. 15, pp. 75–103, 2000.

[9] Instituto Tecnológico de Informática, Fondazione Ugo Bordoni, RWTH Aachen, and ZERES Gmbh, "Final report," 2000.

[10] H. Ney, L. Welling, S. Ortmanns, K. Beulen, and F. Wessel, "The RWTH large vocabulary continuous speech recognition system," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 1998, pp. 853–856.