FILLER MODEL BASED CONFIDENCE MEASURES FOR SPOKEN DIALOGUE SYSTEMS: A CASE STUDY FOR TURKISH

A. Akyol and H. Erdoğan

(Speech and Language Processing Laboratory, Sabancı University, Turkey)

ABSTRACT

Because of the inadequate performance of speech recognition systems, an accurate confidence scoring mechanism should be employed to understand the user requests correctly. To determine a confidence score for a hypothesis, certain confidence features are combined. In this work, the performance of filler-model based confidence features have been investigated. Five types of filler model networks were defined: triphone-network, phone-network, phone-class network, 5-state catch-all model and 3-state catch-all model. First all models were evaluated in a Turkish speech recognition task in terms of their ability to correctly tag (recognition-error or correct) recognition hypotheses. Here, the best performance was obtained from triphone recognition network. Then the performance of reliable combinations of these models were investigated and it was observed that certain combinations of filler models could significantly improve the accuracy of the confidence annotation.

1. INTRODUCTION

For spoken dialogue applications it is critical to understand the user's requests accurately, since the rest of the system acts based on this recognized utterance. On the other hand, the performance of current speech recognition systems are far from perfect. Even when the accuracy is high, robust confidence measures are needed to detect out-of-vocabulary words or non-speech sounds. Thus, an accurate confidence scoring technique should take into account various factors which can contribute to misrecognitions and provide reliable estimates of which words in the output from the recognizer are likely to be correct and which can possibly be incorrect.

A typical approach for confidence scoring includes two steps. First, a confidence feature vector is formed by combining one or more basic features assumed to be correlated with word confidence. Then one or a set of variety of classifiers is applied to this vector to determine the confidence for the recognized word. Quality of the extracted features is the main determinant of the performance of a confidence annotator. So by defining informative features and forming a reliable set of such features, it is possible to increase the performance of the confidence annotation.

In literature, many features have been defined [1,2,3,4,5,6]. They have been grouped with respect to their extraction sources, like acoustic model, language model, semantic, recognition lattice, or n-best list [5]. The use of acoustic features can be very useful for most speech recognition systems. In this work we have investigated the performance change in confidence annotation by defining and combining certain acoustic features based on filler-models with differing levels of acoustic details.

The previous studies showed that the most important confidence feature is the normalized decoder score [3,4,7,8,9]. The confidence of an utterance can be determined by a comparison between the best path from a filler network, and the best path from the regular decoder. Filler model defines a joint HMM that can model any sequence of acoustic sub-units, even when the utterance contains out of vocabulary words. Filler model acts as an acoustic subunit recognizer rather than a word sequence recognizer. Even when the observation sequence is not in the grammar or in the LM, filler model would find a good fit for the input utterance by concatenating the hypothesized sub-units. However the normal decoder tries to find the best match by considering the word models only. It can be concluded that filler models can provide valuable information to detect misrecognitions. In this work we investigate the effect of filler networks in differing detail on confidence estimation. Also combining different filler models could provide better estimates for the hypothesis confidence. Although there could be some overlapping information between the models, there could also be discriminative information specific to each filler model. To make use of all the information in the decision process, we look for the best combination of features obtained from the filler models and other acoustic features. The paper is organized as follows; in section 2 we provide a review of theory and introduce the filler model networks that we used. Then in section 3 we give the details of experiments and in the following section results are reported. Finally we present a summary of this work with a discussion of future work.

2. CONFIDENCE SCORING AND FILLER MODELS

If we assume that an acoustic observation, $O = o_1, o_2, ..., o_t$, is produced by a sequence of words, $W = w_1, w_2, ..., w_n$, then the aim of speech recognition system is to determine the most probable word sequence, $\hat{W}$, given the observed acoustic signal, O , and it is represented by the Bayes' equation (1);

$$\hat{W} = \arg\max_W P(W|O) = \arg\max_W \frac{P(O|W)P(W)}{P(O)} \qquad (1)$$

In most recognisers the denominator, acoustic probability of the observation, is assumed equal for all observations and is not considered in calculations. This means that the recogniser likelihoods are not absolute measures for the probability of O but relative measures used to compare different utterance hypotheses.

We could use the denominator of the fundamental equation (1) to help in computing the confidence information for an hypothesis since the ratio will be an absolute measure of the probability of the word sequence. $P(O)$ can be approximated by general-purpose recognizers based on filler networks. These kind of recognizers should be able to recognize anything, so that they can fill the holes of the grammar in a speech recognizer. In order to do this, the acoustic space is modeled in certain small units and these unit models are connected to each other within a null-grammar (parallel connection and a loop) structure. Thus the recognition process is made free from any effect of constrained networks like a word sequence grammar. In other words, filler models output the

best unconstrained acoustic path from the unit networks, as given in Eq. (2).

$$\begin{aligned} P(O) &= \sum_{U=u_1 \cdots u_N} P(O|U)P(U) \qquad (2) \\ &\approx \arg\max_{U} P(O|U)P(U) \\ &\approx \arg\max_{N, u_1 \cdots u_N} \prod_{i=1}^{N} P(O_{l_i}^{l_{i+1}}|u_i)\ 1/M \end{aligned}$$

Here, U represents all possible unit sequences and $(u_i)_1^N$ denotes N units realized in a sequence U. M is the number of units in the filler network and $O_{l_i}^{l_{i+1}}$ represents the observation sequence that aligns with the unit model u_i. In practice, $P(O)$ in (2) is found using Viterbi decoding on the filler model network.

In general, the confidence of an utterance will be correlated with the Bayes' ratio in (1), which can be approximated by dividing the best path likelihood from the regular decoder approximating $P(O|W)P(W)$ and the best path likelihood from the acoustic unit network approximating $P(O)$ as shown above. The ratio should ideally be less than one. However, some filler networks might result in lower likelihoods than the decoder likelihood depending on how fine/crude the acoustic units in the network are.

As filler model usually all-phone networks or catch-all models are used and the output is used to normalize the decoder's result. In our approach, we propose to use an all-triphone network and a phone-class network in addition to them. It is expected that with the increasing detail used in the unit models, we approximate $P(O)$ better. Thus, we hypothesize that using an all-triphone network would result in better confidence annotation than using an all-phone network. In addition to this we expect significant increase in the performance when such kind of features are combined by an efficient classifier.

2.1. Triphone Recognition Network

In this type of filler network, the acoustic space is modeled by using triphones. Triphone models are powerful since they capture the most important coarticulatory effects. We use Turkish telephony speech data for this work. Our phone set includes 31 phones identified for the Turkish language. To construct a filler model which contains all triphones used in our training data, first, all different triphones were counted. Then, in order to decrease data-overfitting, a unit clustering technique was applied. Thus the number of parameters to estimate was decreased to 1673 triphones (each triphone HMM has 5-state[1]/5-mixture topology) with a generic pause and silence model, as seen in Figure 1.

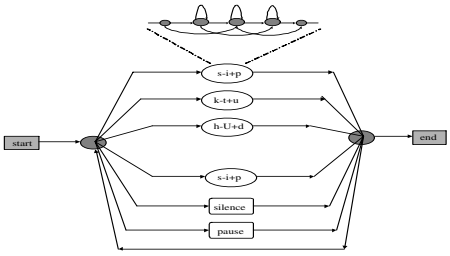

Fig. 1. Triphone-Filler Netwok

[1]Only three of them are emitting states.

The input utterance is recognized with this network as well as the word recognizer and for each word in the recognized sentence, a filler model score is produced by summing all per-frame scores (from the filler model recognizer) aligning with that word.

2.2. Monophone Recognition Network

In this filler model network, we decreased the detail in acoustic modeling and we modeled the whole acoustic space with 31 context-independent phone models, instead of 1673 models, by connecting them with in a null-grammar scheme as similiar to the network of triphones. Now, the number of Gaussians to train is 465, not 25095 as in triphone network.

There are 29 letters in Turkish Alphabet. Since Turkish is almost a *phonetic language*, each letter can be assumed to correspond to a phoneme[2]. That means for each of the letters in the alphabet;
Consonants: b c ç d f g ğ h j k l m n p r s ş t v y z ,
Vowels: a e ı i o ö u ü ,
a phoneme is used. For convenience with English tools we represent the non-English letters with their capital equivalents;
Phonemes: a,b,c,C,d,e,f,g, G,h,I,i,j,k, l,m,n, o,O, p,r,s,S,t,u, U,v,y,z
(ç → C; $\tilde{g}$ → G; ı → I; ö → O; ş → S; ü → U)
In this work, obtaining pronunciation of words is assumed to be a simple one-to-one mapping of letter sequences to phone sequences. In literature Monophone Networks are usually added as alternative paths to the recognizers [1]. But in this work all filler networks are considered as separate decoding processes.

2.3. Phone-Class Recognition Network

Here, we decreased the acoustic detail in the model a bit more. In all-phone filler network, we had used 31 phone models for Turkish Language. Now we cluster these phones into six groups by considering their linguistic characteristics as in Table 1 [11].

Table 1. Phone classes in Turkish used as filler models

Name of Phone Class	Phones Included
Stops	p,t,C,k,b,d,c,g
Fricatives	f,s,S,v,z,j
Nasals	m,n
Liquids	l,r
Glides	y, G, h
Backvowels	a, I, o, u
Frontvowels	e, i, O, U

The models of the groups were trained and they were placed in a fully connected network with silence and pause as in Figure 2.

2.4. Catch-All Models

This filler model is the simplest and the widely used one among the filler models [1, 2]. The idea is to model all acoustic variability inside only one model. In this work two different topologies (different number of states and different number of mixtures) were tried. In the first one 5-state topology was used and for the other one we used 3-state structure, as seen in Figure 3.

3. EXPERIMENTAL SETUP

We experimented the proposed confidence annotators on Sabancı University Automated Course Inquiry System. Briefly, the sys-

[2]As always, there are exceptions.

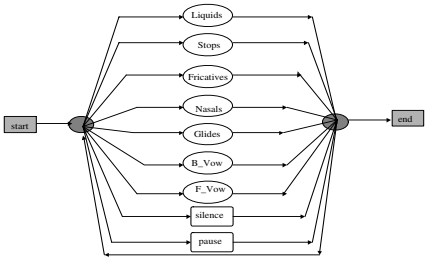

Fig. 2. Phone Class Filler Netwok

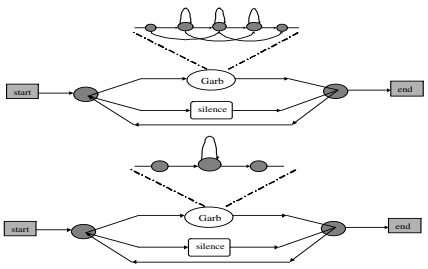

Fig. 3. Catch-All Filler Netwok, 5-state topology and 3-state topology

tem accepts calls from students and enables them to ask questions about classes, such as the lecturer, the time, the place, the number of credits and etc. This spoken dialogue system uses triphone HMMs and works on a pre-defined grammar.

To train the filler models we used two separate databases. First one was a general-purpose Turkish speech database, TurTel [11], collected over the public telephone network by using three types of microphone; handset, hands-free, and cellular phones. Its corpus is based on statistical triphone modeling of Turkish and it assumes that %80 of Turkish can be represented by 1000 triphones. These triphopnes were covered by 15 sentences and 373 words in this database. The speaker set consists of 57 male and 36 female from different age groups and origins in Turkey . The other one was an application-specific database collected for this course inquiry system, so that all utterances are inquiries about courses. It consists of 4500 utterances from 45 speakers. But only %50 of this database was used, together with TurTel, for model training. %30 of this data was used for classifier training and the remaining %20 was used in testing the confidence annotation system.

It is hard to measure and compare performances of confidence annotators, since the performance of such systems depends on application specific parameters. In this work, we consider confidence annotation as a two-class problem where we only mark a recognized word as *correctly recognized* or *recognition error*. In the literature different criteria for the evaluation of confidence measuring systems have been proposed like EER, CER, NCE, NERP and so on [1,6,12,13]. In this work EER (Equal Error Rate) is used. It is the operating point of the classifier where the False-Accept (FA) rate is equal to the False-Reject (FR) rate. In this point, the receiver operating characteristic (ROC) curve is closest to the origin of the FA and FR axes. The ROC curves are used to represent the performance of a confidence measure and they intersect the axes of FA and FR and plot the operating point on the FA-FR plane by varying the decision (or classification) threshold.

In our experiments we extracted 12 candidate features for each hypothesized word. All of them were acoustic and word-level features and most of them(8 of 12) were extracted from the parallel filler model decodings using the hypothesized word boundaries .

1. Per-Frame log-likelihood Score (LL)
2. Per-Frame log-likelihood Score of the triphone network (TL)
3. Per-Frame log-likelihood of the monophone network (PL)
4. Per-Frame log-likelihood Score of the phone-class network(CL)
5. Per-Frame log-likelihood of the 5-state catch-all model (CF)
6. Per-Frame log-likelihood of the 3-state catch-all model (CT)
7. Triphone log-likelihood ratio score (TR)
8. Monophone log-likelihood ratio score (PR)
9. Maximum frame score (MA)
10. Minimum frame score (MI)
11. Standart deviation of frame scores in the hypothesis (SD)
12. Number of phones in the hypothesis (NP)

Here, since we derived frame level scores, for all score calculations we used normalized unit scores and so that we called them *per-frame* scores. Also ratio scores (feature TR and feature PR) were calculated by dividing the *per-frame filler network score* to the *per-frame normal decoding score.*

4. RESULTS

First the individual performance of certain filler model network features were investigated to give sense about the quality of the filler networks. So 5 representative features were selected for each filler network and their EER values were calculated as in Table 2. We use a simple GMM classifier. GMM classifier models the features in each class with a mixture of Gaussians. The Gaussian mixtures are trained using classifier training data which contains recognized words labeled as *correct* or *incorrect*. For testing, likelihood ratio with a varying decision threshold is used. Here we want to note that for each EER value calculation, 40 different mixture combinations were tried, and for each mixture combination, 140 different threshold values were experimented.

Table 2. Individual filler network performance

Filler Network Type	EER(%)
Triphone Recognition Ratio - (TR)	27.12
Monophone Recognition Ratio - (PR)	33.49
Phone-Class Recog. Net - (LL,CL)	38.78
5-state Catch-All Model - (LL, CF)	38.37
3-state Catch-All Model - (LL, CT)	42.78

According to table 2, it can be concluded that with more detailed acoustic modeling in the filler network, better results, namely lower EERs, could be obtained. Also despite the decrease in individual performance, each filler model type could provide different information about the confidence of the recognition. In other words, all filler model scores may contain overlapping information but on the other hand they may also contain some individual information, which can not be obtained from the other filler model types.

In Table 3 we present the performance of different combination schemes to understand the trade-off between the overlapping and discriminating confidence information of filler model networks in different acoustic details. (Actually we investigated 60 different

Table 3. Feature combination results for GMM classifier. Feature codes are given in Section 6.2

Included Features	EER(%)
TR	27.82
PR	33.49
LL,CL	38.78
LL,CT	42.78
LL,TR	26.98
LL,PR	34.42
TR,PR	25.62
LL,TR,PR	26.52
LL,MA,MI	37.03
TR,PR,NP	24.32
CL,TR,PR	25.35
LL,TR,PR,SD	25.55
CL,CF,TR,PR	26.23
TR,PR,MA,MI	26.43
LL,CL,CF,TR,PR	28.57
CF,TR,PR,SD,NP	25.15
CL,CF,TR,PR,MA,MI	27.64
CL,CF,TR,PR,SD,NP	26.73
CL,CF,TR,PR,MA,MI,NP	27.43

combinations, but in table 3 we only provide significant 19 of them.) The Feature TR, *triphone log-likelihood ratio*, provides the best EER, % 27.82, among the individual performance. The difference between the best result in the table, Combination (TR,PR,NP): % 24.32, is only %3.50, and also it is included by the best combination. Then it is reasonable to say that the most of the discriminating information in the best feature combination comes from this feature. It is an expected result since the feature TR uses the information of the most detailed modeling of unconstrained speech. The substantial contribution of Feature TR can be also seen from the good performance of the other combinations which include TR.

Also the % 3.5 performance increase between the Combination (TR), a single feature combination, the EER is %27.82 and the best EER combination (TR,PR,NP) in the table, %24.32, is significant if we think that the cost of extracting these extra features is much less than the one of the baseline feature, TR. For triphone log-likelihood ratio score, feature TR, we search on a network of 1675 models but for phone log-likelihood ratio score, Feature PR, we search on a network of only 31 models.

Although the Combination (CL,CF,TR,PR,SD,NP) includes the best combination (TR,PR,NP), its performance (% 26.73) is less than the performance of its subset, (% 24.32). This means that this combination includes features that do not help in classification that can be called noise and this noise badly effects the overall usefulness of the feature vector. Another reason for this case could be the curse of dimensionality problem. It is possible that the classifier training data is insufficient to determine a boundary on a 6-dimensional space.

5. CONCLUSION AND FUTURE WORK

We investigated the performance of filler models in different details for grammar-based continuous speech recognition systems. After the evaluation of features in terms of their ability to increase the confidence annotation accuracy, we investigated the performance improvement when they were combined in the same feature vector.

Among the filler model types investigated, the best individual performance was obtained from triphone recognition network with a significant improvement, %5.67 as compared to the all-phone network, a popular filler model type in literature. The main reason of the improvement is that the triphone network uses much more detailed acoustic models than the other filler model types. Although this detailed modeling brings some problems like trainability and computational cost, these disadvantages can be alleviated by some implementation tricks like parameter-tying and pruning.

In general, one type of filler model is used to define a reliable confidence information for the hypothesis. But it was observed that combinations of different filler model types increased the confidence annotation performance, %3.5 as compared to the best filler model performance, triphone recognition network.

This work has opened up many new research possibilities on increasing the confidence annotation performance. First, alternative filler model types can be constructed like all-syllable network or null-grammar uni-gram network. Since these topologies also provide detailed modeling of the acoustic data, improvement on the determination of the confidence on the hypothesis can be expected. Also the performances of reliable combinations of filler models and other types of features can be investigated. Lastly, to increase the accuracy in Turkish speech recognition systems, using a more detailed phone-set could be useful. Oflazer et al. [10] proposed a new phone-set for Turkish, which consists of 45 different phones instead of standard 29 phone definitions.

6. REFERENCES

[1] Chase L. Error-Responsive Feedback Mechanisms for Speech Recognition. Phd Thesis, Carnegie Mellon University, 1997.
[2] Hazen T.J., Burianek T., Polifroni J., Seneff S. Recognition confidence scoring for use in speech understanding systems. Computer Speech and Language16, 49-67, 2002.
[3] Cox S., Dasmahapatra S. A high-level approach to confidence estimation in speech recognition. Eurospeech 1999.
[4] Kamppari S., Hazen T., 2000. Word and phone level acoustic confidence scoring. ICASSP 2000.
[5] Zhang R., Rudnicky A.I., Word level confidence annotation using combinations of features.Eurospeech 2001.
[6] Schaaf T., Kemp T. Confidence measures for spontaneous speech recognition. ICASSP 1997.
[7] Gunawardana A., Hon H. W., Jiang L. Word-based acoustic confidence measures for LVSR. ICASSP 1998.
[8] Weintraub, M. LVCSR log-likelihood ratio scoring for keyword spotting. ICASSP 1995.
[9] Cox S., Rose R. Confidence measures for the switchboard database. ICASSP 1996.
[10] Oflazer K., Inkelas S. A Finite State Pronunciation Lexicon for Turkish. EACL Workshop on FSMs in NLP 2003.
[11] Yapanel U. Garbage modelling techniques for a Turkish keyword spotting system, Msc. Thesis, Boğaziçi University, 2000.
[12] Maison B., Gopinath R. Robust confidence annotation and rejection for continuous speech recognition. ICASSP 2001.
[13] Weintraub M., Beaufays F., Rivlin Z., Konig Y., Stolcke A. neural-network based measure of confidence for word recognition.