Proc. of the National Conference on Communications (NCC), Feb 2013, IIT Delhi, India

Confidence Measures for Detecting Speech Recognition Errors

Jigar Gada, Preeti Rao
Department of Electrical Engineering,
Indian Institute of Technology Bombay,
Mumbai 400076, India
{[1]jigar23, [2]prao}@ee.iitb.ac.in

Samudravijaya K
School of Technology and Computer Science
Tata Institute of Fundamental Research
Mumbai 400005, India
[3]samudravijaya@gmail.com

Abstract—**Errors of speech recognition systems occur due to a variety of reasons. It is desirable to have a confidence measure that gives an idea of the accuracy of the decoder output, so that appropriate remedial measures can be taken. In this paper, we compare two approaches to detect incorrect output of a speech recognition system. The first approach employs multiple decoders, and uses a voting method to surmise confidence in the accuracy of the speech recognition system. The second approach uses a single decoder, but judiciously combines information at the segmental as well as supra segmental level to derive a measure of confidence in the output of the decoder. A neural network is trained with three features based on phone duration and one feature based on acoustic score. The output of the neural network is used to estimate the confidence in the output of the decoder. The two approaches are compared for their efficacy in detecting utterances that do not contain a valid input according to the task grammar as well as wrongly recognized valid inputs. It was observed that the second method achieves much better rejection of invalid input utterances as compared to the multi-decoder method, despite decoding a test utterance just once.**

Keywords: **Automatic Speech Recognition, Out of Vocabulary, Confidence Measures, CMU-Sphinx.**

I. INTRODUCTION

Speech-based human-machine interaction technology has gained popularity over the last decade due to its important commercial implications. Depending on the complexity of the task and the variability in the input speech, Automatic Speech Recognition (ASR) systems have met with varying degrees of success. To enhance the practical utility of such imperfect systems, it would help to incorporate a measure of confidence in the speech recognizer output within the human-machine dialog framework. For example, it would be useful to know when a particular instance of decoded output is unreliable so that suitable confirmation queries can be raised before proceeding with the dialog. Unreliable outputs can arise for a number of reasons, viz. noisy or degraded speech input, speaker accent leading to mismatch with respect to training data, and out-of-vocabulary utterances.

A confidence measure (CM) can be generated for each recognition output based on additional information that is usually available as a by-product of the decoding process. We consider the issue of confidence measures in the context of an on-going project on telephone speech based access of agricultural commodity prices in markets of Maharashtra.

The speech database comprises of the names of 31 districts, 279 markets and over 300 unique commodities. The system downloads, on a daily basis, prices of agricultural commodities from a website (http://agmarknet.nic.in/) maintained by the Ministry of Agriculture, Government of India. Due to noisy/degraded speech as well as unexpected inputs from farmers, the accuracy of the ASR system is not as high as desired. In order to improve the accuracy of the system, a hierarchical querying framework is assumed with a limited vocabulary at each node. As illustrated in Fig. 1, each district has a set of *mandis* (markets) which, in turn, have a set of commodities traded in each mandi. This structure helps in reducing the number of active words in the vocabulary at a given instant without compromising the accuracy.

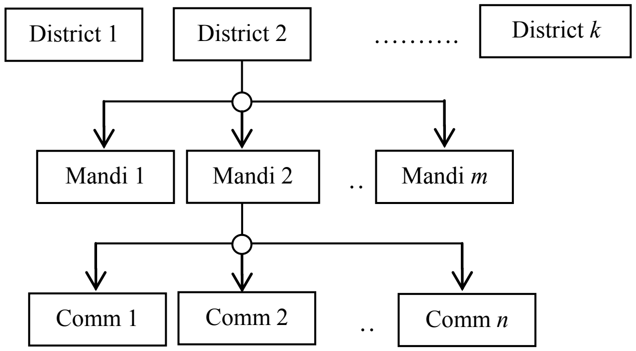

Figure 1. Division of vocabulary into different sets in a hierarchical querying framework.

A hierarchical structure implies engaging the user in a dialog by asking questions at each node. An example of a dialog or call-flow is shown in Table I.

An easy approach to deal with recognition errors is to explicitly take account the fact that the decoder output may not be correct every time. To avoid entering a wrong node, and taking the dialog along an irrecoverable path, an explicit step of confirmation by the caller about the validity of the decoder output is built-in at every stage, as shown in Fig. 2.

978-1-4673-5952-8/13/$31.00 ©2013 IEEE

TABLE I. Sample conversation for speech based access of agricultural commodity prices.

System: Please speak the name of the district
User: *Ahmadnagar*
System: Please speak the name of the mandi
User: *Jamkhed*
System: Please speak the name of the commodity
User: *kaapus*
System: The price of the *kaapus* is …

However, adding an explicit confirmation at every node not only increases the call time but also annoys the user. Therefore, we seek an objective measure of reliability of the decoder output, or a confidence measure (CM), based on which the confirmation step can be applied selectively. While a clearly articulated utterance of an in-vocabulary word is expected to elicit correct recognition with a high CM, a noisy utterance or an Out Of Vocabulary (OOV) word would result in a low CM. If the CM is high, the confirmation step can be skipped altogether. If the CM is in the medium range, the system seeks, from the user, a confirmation of the decoder output. On the other hand, if CM is low, the system immediately requests the user to repeat the utterance.

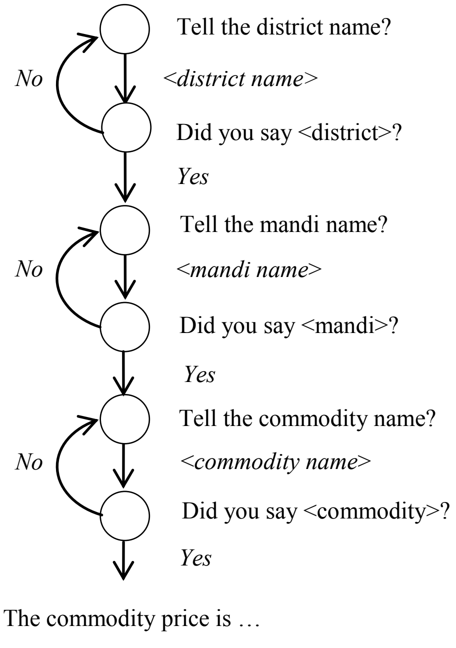

Figure 2. Conversation flowchart with mandatory user confirmation at each node of the dialog tree.

A number of methods of confidence scoring have been reported in the literature [1]. These include voting based on multiple decoder outputs [2], confidence measures based on phone duration [3] or log-likelihood [4] etc. This paper describes a measure of confidence derived from the output of a single decoder, and compares its efficacy in detecting utterances that do not contain a valid input according to the task grammar. The test data comprises a significant amount of noisy and disfluent speech compared with the previously reported works [3, 4].

II. DATA COLLECTION AND ACOUSTIC MODELS

An interactive voice response system over telephone channel was used to record in-vocabulary words uttered by farmers following 3 approaches [8]: (i) Users are asked selected questions and the answers are recorded, (ii) Users are given a list of district, mandi and commodity names, and are asked to read them, (iii) Users are shown pictures of commodities, and are asked to utter their names; this approach provided pronunciation variants for the commodity names.

Since the system is being developed for use by farmers, majority of the data is collected from rural areas of Maharashtra. The database used for training subword acoustic models consists of about 45,000 utterances from 1500 native speakers. This data, labeled at the word (including fillers) level by human transcribers, was used to train 68 phone models and 11 filler models using the CMU-Sphinx tool [7]. These are context dependent; tied-triphone HMM models with number of senones in the range 1000-2000 and the number of Gaussian mixtures per senone in the range 8-16.

III. CONFIDENCE MEASURES

Even though the user is expected to say just one of the names the system is expecting, the recorded speech contains quite a few types of extraneous sounds. These can be categorized as (a) OOV words: e.g., user speaks a mandi name when the system expects a district name. (b) Speech disfluencies: human generated sounds such as ‘aah’, cough. (c) background noise including speech of other speakers (babble). Such unwanted sounds are in conflict with the FSG, and lead to recognition errors. The role of a confidence measure is to detect such errors, and seek feedback from the user when the confidence is not high. Section A describes a measure of confidence derived from the output of multiple speech decoders. Section B describes the proposed measure of confidence that is derived from supplementary outputs of a single decoder.

A. Using Multiple Decoders

In this method, we use three decoders which are trained with different parameters. The parameters which vary for these decoders are the number of senones and the number of mixtures in the Gaussian Mixture Model of a senone [6]. The strategy of exploiting outputs of multiple decoders is similar to that of [2], but with a major difference. All the 3 decoders used in the current work hypothesize sentences in a given acoustic feature vector space, but were trained with different configuration parameters of HMMs. In contrast, the decoders of [2] operate in different acoustic feature vector spaces.

In the current work, the 3 decoders listed in Table II were chosen due to their high word recognition accuracies in 3-fold cross validation experiments.

TABLE II. Specifications and recognition accuracies of the three decoders used in multiple-decoder based ASR system.

	No. Of Senones	*No. Of Gaussians*	*Avg word accuracy(%)*
Decoder 1	1000	16	70.2
Decoder 2	2000	8	68.7
Decoder 3	2000	16	67.7

Let d1, d2 and d3 denote the 3 decoders; let h1, h2 and h3 denote the corresponding sentence hypotheses for a test utterance. We consider 3 cases for confidence measures.

Case 1: h1 = h2 = h3 [***high confidence***]

Action: Thanks to unanimity among the 3 decoders, the system jumps directly to the next node of the dialog tree without asking for confirmation from the user.

Case 2: Any two decoder outputs are same, e.g., h1 = h2 ≠ h3 [***medium confidence***]

Action: The system orally communicates h1 to the user and seeks a confirmation. For example, if the output of 2 decoders is *kolhapur* and output of the other decoder is *solapur* then user will be asked 'I heard *kolhapur*. Is it correct or wrong?' If the user response is 'correct', the system proceeds to the next node; otherwise the system remains in the current node, and asks for the name again.

Case 3: h1 ≠ h2 ≠ h3 [***low confidence]***

Action: The user will be told 'Sorry I didn't get you'. The system remains in the current node, and seeks information again.

However, this method is computationally expensive as every test utterance has to be decoded by each of the 3 decoders. It is desirable to have a better method of estimating the accuracy of decoder output that does not increase computational load, and yet is effective in reducing the number of explicit user confirmations. Section B proposes such a scheme based on outputs of a single decoder.

B. Using Single Decoder

In the proposed approach to minimize the number of confirmation steps in the spoken dialog, a test utterance is decoded only once. The next action of the system is decided based on the value of a CM that depends not only on the likelihood of the sentence hypothesis, but also on 3 suprasegmental features computed from auxillary output of the decoder. The decoder 1 (1000 senones and 16 mixtures) of Table II is selected due to its superior recognition accuracy. CMU-Sphinx based aligner is used to force-align the test utterance with the Viterbi decoder hypothesis. This yields segmentation of the utterance in terms of phones. Four features are derived from the acoustic likelihoods as well as durations of the phone segments. The training data is used to train a neural network to output high value when fed with 4 features of correct decoder hypothesis, and low value in case of incorrect decoder hypothesis. The confidence measure of a test utterance is a function of the output of the neural network.

1) Phone Duration Based Features

In case of incorrectly decoded utterances, we observed that a significant number of phone segment durations are abnormally high. The first 2 features exploit this fact. For each phone, parameters (mean and SD) of a Normal distribution were estimated from the training data. Let *n99* and *n95* denote the no. of phones (in the hypothesis) whose durations are longer than the 99 and 95 percentile respectively [3]. In other words, *n95* is the no. of phones whose durations are greater than mean+1.96SD of their respective duration distributions. The first two features are computed as *n99/N* and *n95/N* where *N* is the total number of phones in the hypothesis. These features are an estimate of the fraction of phones with abnormal durations.

The 3rd feature used was an estimate of the speaking rate computed as follows [3].

$$\alpha = \frac{1}{N}\sum_{i=1}^{N}\frac{d_i}{x_p} \qquad (1)$$

where d_i denotes the duration of the i^{th} phone segment in the utterance, x_p is mean length of the corresponding phone (learned during training), N is the total number of phones in the utterance.

2) Acoustic Score Based Feature

In addition to the 3 duration-based features described above, we used an additional feature that is related to the acoustic score. The output of the Sphinx decoder provides log likelihoods of every acoustic segment in the best hypothesis. We define *acoustic score* of a phone segment as the ratio of the log likelihood of the segment divided by the duration of the phone segment.

Inspection of segment acoustic scores of different utterances showed that the segment acoustic scores of phones for most of the out of vocabulary utterances were less than -20000. Let n denote the number of phones in the utterance with acoustic score smaller than -20,000. The fourth feature is computed as *n/N* where *N* is the total number of phones in the utterance.

3) Detection Of Recognition Error

An Artificial Neural Net (ANN) was used to classify the decoder output in a post processing step [3]. A 3-layer neural network was trained with 3 prosodic features (described in Section 1), and an acoustic score based feature (described in Section 2). The ANN consists of 4 nodes in the input layer, 9 nodes in the hidden layer and a single output node. Experiments were conducted at the district node (see Fig. 2) having 1 FSG of 31 districts, mandi node which has 31 FSGs with each FSG having, on an average, 12-20 mandis and the commodity node which has around 320 FSGs with each FSG having on an average 50 commodities. A neural network is trained for each of the 3 nodes. The results are reported here for the district node only; the other nodes showed comparable performances.

During training, features of utterances containing a district name, with (noisy files) or without (clean files) speech disfluencies, that were *correctly recognized by the decoder* were fed with +1 as the desired output of the ANN. Negative exemplars (where desired output of ANN is -1) were utterances containing (i) just speech disfluency (babble, <hm>, <aah>, <laugh> etc.) referred to as Voiced filler (ii) Silence and background noise (<horn>, <bang> etc.) referred to as Silence_uv, and (iii) OOV words (specifically commodity names). Given this, the output of the trained ANN for a test utterance will be in the range [-1 1].

Positive data (files which are to be accepted) comprised of 913 clean district files and 368 noisy district files that have been correctly decoded by decoder 1. Negative data (files which are to be rejected) comprised of 750 silence_uv files, 990 files with Voiced fillers and 600 commodity/OOV files. Number of files used as negative exemplars have been chosen based on some experiments performed by varying the number of files, and selecting the best set. A 3-fold cross validation experiment is performed with 2/3rd data used for training and remaining 1/3rd data used for testing.

The system employing multiple decoders classifies the decision associated with a decoded test utterance into 3 classes: High, Medium and Low confidence. For comparison purposes, the output of the ANN was used to classify a decoded test utterance into the same 3 classes using two thresholds, set empirically. The thresholds can be fixed depending on the desired trade-off between the acceptance of correct decoding (to be maximised) and that of wrong decoding (to be minimised). In our case, threshold of -0.4 and 0.1 proved to be effective in categorizing the output of the ANN (in the range [-1, 1]). For a test utterance, if the value at the output node of ANN is greater than 0.1, the system will jump directly to the next node (High confidence). If the output is between -0.4 and 0.1, the system will ask the user for a confirmation of the decoder output (Medium confidence). If the output is less than -0.4, the system will ask the user to repeat the utterance (Low confidence).

IV. EXPERIMENTAL RESULTS AND DISCUSSION

In this section, we present and compare performances of CM systems using the two approaches for the same test data.

The test files can be categorized into 3 groups depending on (a) the content of input wave file and (b) the output of the single decoder used in the second approach. If a wave file contains a valid input (the name of a district), we refer to such files as 'district files'. If such a district file contains any speech disfluency or extraneous sounds, we call it a 'noisy district file'. The output of the decoder when fed with a district file may agree with the ground truth (correct recognition) or may not (incorrect recognition). The value of an ideal CM should be high in case of files that contain a district name, and the utterance was recognized correctly by the decoder. On the other hand, the CM should be low for incorrectly decoded district files. In addition, the CM should be low for files that do not contain valid speech (a district name). Such an invalid input file could belong to one of the 3 types: (i) does not contain any speech (ii) contains only speech disfluencies or babble, (iii) contains an Out Of Vocabulary word.

A. Experimental Results

For the sake of brevity, let us denote the two systems following the two approaches as follows:

M1: Multiple decoder output based Confidence Measure.
M2: Neural Network based Confidence Measure.

Tables III, IV and V show the distribution of test utterances of various categories in terms of the confidence classes. Table III shows the CM classes of two methods for 'district files' that contain name of a district and the output of the decoder is correct. Table IV shows the CM classes of the two methods when input files do not contain valid input. Table V shows the corresponding figures in the cases when the input file contained a district name, but the decoder output was wrong.

TABLE III. Confidence estimated by two systems for clean and noisy district files that were correctly decoded.

	Clean district (%)		Noisy district (%)	
	M1	*M2*	*M1*	*M2*
High	98	80.1	82.9	56.5
Medium	1.9	7.5	12.5	13.1
Low	0.1	11.1	4.6	28.5

TABLE IV. Confidence estimated for silence_uv, voiced fillers and OOV/commodity files.

	Silence_uv (%)		Voiced filler(%)		OOV words (%)	
	M1	*M2*	*M1*	*M2*	*M1*	*M2*
High	34.6	4.4	40.3	9.2	33.9	8.3
Medium	46.1	4.9	44.4	9.9	41.2	9.5
Low	19.3	90.5	15.3	80.3	13.3	80.7

TABLE V. Confidence estimated for clean and noisy district files that are incorrectly decoded.

	Clean district (%)		Noisy district (%)	
	M1	*M2*	*M1*	*M2*
High	33.9	36.3	82.8	67.8
Medium	55.4	11.3	14.9	9.6
Low	10.7	49.4	2.3	19.5

B. Discussion

Ideally, we would like all correctly decoded files (Table III) to lie in the High confidence class, and all other categories (Tables IV, V) to lie in the Low confidence class.

It is desirable that the percentage of utterances that do not contain a valid word [(a) silence_uv (b) voiced fillers, (c) commodity/OOV], but are accepted with high confidence by the system should be very low. Otherwise, the system goes irrecoverably into a wrong branch of the dialog tree. According to row 1 of Table IV, the percentage of such serious error is in the range [34%-40%] in case of multiple decoders based CM; it reduces to [4%-9%] in case of ANN based CM. Since the reduction in such serious errors is by a factor of 4.5 to 8, one can conclude that the CM based on ANN has achieved the prime objective reasonably well. In addition, the percentage of low confidence cases for the above 3 types of files (refer last row of Table IV) is much higher [80%-90%] for ANN based CM than for the multiple decoder system [13%-19%]. This, in turn, significantly reduces the number of times confirmation is sought from the farmer (a time consuming/annoying situation). On the other hand, the penalty to pay for such gain is some increase in the percentage of low confidence cases when the decoder recognized name correctly. This (refer last row of Table III) increased from 0.1% to 11.1% in case of clean district name files, and from 4.6% to 28.5% in case of district name files that also contain fillers. Thus, the rate of False Acceptance is significantly reduced accompanied by some increase in False Rejection rate. A redeeming feature of the ANN based CM is that it is able to detect and reject nearly half of the incorrect decoder outputs in case of clean utterances (refer last row of Table V). In contrast, the multiple decoder system is able to detect just 10.7% of such cases.

For the ANN based CM, errors tend to occur more for phones like /a/, /aa/, /ii/ where the variance of phone duration is quite high making duration modeling more challenging. Further, many silence/filler utterances tend to be decoded to short words (3 to 5 phones). During alignment, often one of the phones is assigned an abnormally high duration (> 100 frames) while other phone durations are normal; the latter results in high CM. Such cases can be handled by assigning low confidence to utterances associated with abnormally high phone durations.

V. CONCLUSION

In this paper, we compared two methods of estimating confidence in the decision of an ASR system. The first method decodes a test utterance thrice using 3 different decoders; the degree of agreement among the outputs of the decoders is taken as the confidence measure. The second approach recognizes a test utterance using a single decoder. It derives a measure of confidence as a function of the output of a trained neural network fed with supplementary information of the decoder. The second approach not only reduces the response time, but also detects utterances that do not contain a valid input more efficiently. Moreover, errors of the decoder in case of noisy files containing valid input are better detected by the second approach. Thus, the second approach that uses information at both segmental and supra segmental levels seem to be more suitable. To increase the true acceptance rate, a study is being carried out on new features as well as on efficient combination of the two approaches.

ACKNOWLEDGMENT

This work was supported by the project ‘Speech-based access of agricultural commodity prices in six Indian languages’ sponsored by Ministry of Communication and Information Technology, Government of India. The authors would like express their gratitude to Dr.Nandini Bondale, Nikul Prajapati, Pranav Jawale, Pinki Kumawat, Tejas Godambe, Namrata Karkera and Vishal Khadake for their support and assistance in the project work.

REFERENCES

[1] Hui Jiang, Confidence measures for speech recognition: A survey, Speech communication, Volume 45, pages 455-470.

[2] Gautam Varma Mantena, S. Rajendran, Suryakanth V. Gangashetty, Kishore Prahallad, “Development of a Spoken Dialogue System for accessing Agricultural Information in Telugu”, Proceedings of ICON-2011, 9th International Conference on Natural Language Processing.

[3] Silke Goronzy, Krzysztof Marasek, Ralf Kompe, Andreas Haag, “Phone-Duration-Based Confidence Measures For Embedded Applications”,*Proceedings of Sixth International Conference on Spoken Language Processing*, ICSLP 2000 / INTERSPEECH 2000, Beijing, China, October 16-20, 2000.

[4] Joel Pinto, R.N.V. Sitaram, “Confidence Measures in Speech Recognition based on Probability Distribution of Likelihoods”, *Proceedings of Interspeech 2005 – Eurospeech 9th European Conference on Speech Communication and Technology*, Lisbon, Portugal, September 4-8,2005.

[5] Daniel Jurafsky & James H. Martin, *Speech and Language processing,* Upper Saddle River, New Jersey: Pearson, 2009, pp. 285-333.

[6] Mei-Yuh Hwang, Xuedong Huang, “Sub-phonetic Modelling for Speech Processing”. HLT’91 *Proceedings of the workshop on Speech and Natural Language*, Pages 174-179.

[7] CMU-Sphinx: Open Source Toolkit for Speech Recognition. http://cmusphinx.sourceforge.net/.

[8] T. Godambe and K. Samudravijaya, “Speech data acquisition for voice based agricultural information retrieval”, Proc. Of 39th All India DLA Conference, Punjabi University, Patiala, June 2011 <http://speech.tifr.res.in/chief/publ/11DLA_agriSpeechDataAcquisition.pdf>.