

Confidence Measures in HMM/MLP Hybrid Speech Recognition for Turkish Language

Erhan Mengusoglu, Olivier Deroo

*Multitel-TCTS
Parc Initialis, Avenue Copernic,
7000 Mons, Belgium*

E-Mail: {mengus,deroo}@tcts.fpms.ac.be

Abstract— A confidence measure is defined as the posterior probability of word correctness given the values of confidence indicators [8]. Confidence measures can be calculated from a posteriori probability of recognized word or sub-word unit inferred from some acoustic models and language models, including various normalization techniques. In this work we present several confidence measures and propose to apply them to automatic recognition of the Turkish language.

Turkish language is an inflected language. It is possible to produce a very high number of words from the same root with suffixes [4]. Hence confidence measures appear to be promising techniques to improve the current performance of the speech recognition for Turkish language.

For experimental purpose, a Turkish database has been developed. Its content will be described in this communication. The Speech Training and Recognition Unified Tool (STRUT) toolkit has been used for training of the models used in the speech recognition system.

Keywords— Confidence measure, Turkish speech recognition, Language modelling for agglutinative languages, Hybrid speech recognition, Phonetic labeling.

I. INTRODUCTION

A typical HMM based speech recognition system tries to model the speech. The probability of a model given input speech ($P(M|X)$, where M is the model and X is a sequence of acoustic vectors representing a portion of speech) must be calculated. This probability can be expressed in terms of the contributions of an acoustic model and a language model.

Acoustic models are based on speech signal. The first step in acoustic modeling of speech is to determine the model parameters. First order left to right HMM models with self loops are generally used for acoustic models. If HMM emission probabilities are estimated with an Artificial Neural Network (ANN) such as Multi-Layer Perceptron (MLP) the resulting system is HMM/ANN (MLP) Hybrid Speech Recognition System.

Language models bring higher level information by encoding the syntax of the language. To create a language model we need a large text including (nearly) all possible word sequences in the language.

The acoustic and language models will be used to determine the most probable word sequence for a given utterance.

To improve the accuracy of speech recognizer, there is a need for a measure of confidence for recognized speech. If we can define such a measure, the reliability of the system will be increased. Confidence measures can be calculated on both acoustic and language models. Some confidence measures are based only on acoustic models. They use posterior probabilities of HMM decoding. Other confidence measures use language model probabilities. There are some confidence measures that use both of them.

In this paper, the summary of confidence measures used so far, the description of Turkish morphology, the information about database preparation for Turkish language and the description of the use of the Speech Training and Recognition Unified Tool (STRUT) to build a speech recognition system for Turkish will be explained. We will give some of the first results achieved with our baseline system on different tasks.

II. CONFIDENCE MEASURES

A confidence measure is defined as the posterior probability of word correctness given the values of confidence indicators. It can be calculated from posterior probability of recognized words.

Confidence measures are generally based on scaling of acoustic model likelihoods by the likelihood of an alternate model or on training of an application specific post-classifier [8]. In the case of likelihood ratio an alternate model is trained using discriminative training techniques and the likelihood of this model is used to scale the likelihood of the original model. Post classifiers can be based on likelihood of word sequences from the recognizer, language model and some acoustics like speaking rate, signal to noise ratio (SNR).

The likelihood ratios can be used as a test statistic in two different ways; Frequentist and Bayesian. The first one defines a critical region of:

$$\frac{P(x|H1)}{P(x|H0)} \geq T$$

for some non-negative constant T , $H0$ is the hypothesis to be tested, likelihood from recognizer, and $H1$ is the testing hypothesis, likelihood from alternate model.

In the case of Bayesian, the likelihood of hypothesis, $P(x|H)$, is scaled with $P(x)$ and the result is taken as test statistic.

Confidence measures are generally used for utterance verification, keyword spotting and Out Of Vocabulary (OOV) word spotting tasks.

There are three types of confidence measure used; Acoustic measures, Grammatical measures and Combined measures.

The baseline for acoustic confidence measures is local phone class posterior probabilities which is provided by the acceptor HMM. Acceptor HMM provides an estimate of the probability that the model accepted the acoustic. For an acoustic model, acceptor HMM estimates posterior probability of this model given the observations. HMM/MLP hybrid models use acceptor HMMs. Another type of HMM is the generative HMM which provides an estimate of the probability that model generated the acoustic. In the case of generative HMM, the likelihood of the observations given an acoustic model is provided.

The confidence measures defined below are mainly extracted from [8], [7] and [1].

There are four acoustic confidence measures.

Normalized Posterior Probability: This confidence measure is calculated by duration normalization of the local posterior probabilities which is obtained from the acceptor HMM. It can be formulated as:

$$nPP(q_k) = \frac{1}{D} \sum_{n=n_s}^{n_e} \log(P(q_k^n|X_1^n)) \quad (1)$$

where n_s and n_e are the beginning and ending frame of the current phone and D is the duration ($D = n_e - n_s + 1$). $P(q_k^n|X_1^n)$ is the probability of phone class k given the acoustic vector.

This measure gives a score for decoded word or sub-word unit. The result can be used for accepting or rejecting the corresponding decoded sentence. If the result is below a defined threshold it can be said that the confidence for this decoding is low or it is an OOV word.

Scaled Likelihood: This measure can be obtained by dividing the local posterior probability estimate by the class prior probability obtained from the labeling of acoustic training data. This can be formulated as;

$$\begin{aligned} nSL(q_k) &= \frac{1}{D} \sum_{n=n_s}^{n_e} \log\left(\frac{P(q_k^n|X_1^n)}{P(q_k)}\right) \quad (2) \\ &= nPP(q_k) - \log(P(q_k)) \quad (3) \end{aligned}$$

Online Garbage: To calculate this measure, for each frame the average of the m -best likelihood is taken, then this average is normalized over a specified duration (word or sub-word) and this result is used to normalize scaled likelihood confidence measure defined above. It can be formulated as;

$$\begin{aligned} nOLG(q_k) &= \\ \frac{1}{D} \left[SL(q_k) - \sum_{n=n_s}^{n_e} \log\left\{ \frac{1}{m} \sum_{l=1}^m \frac{P(q_l^n|X_1^n)}{P(q_l)} \right\} \right] \quad (4) \end{aligned}$$

Per-Frame Entropy: Unlike the three measures above this measure does not need a decoding (Viterbi state sequence) and can be calculated before the search for best decoding. This measure is the per-frame entropy of the K local phone probabilities averaged over a selected interval. It is calculated by;

$$\begin{aligned} S(n_s, n_e) &= \\ -\frac{1}{D} \sum_{n=n_s}^{n_e} \sum_{k=1}^K P(q_k^n|X_1^n) \log\{P(q_k^n|X_1^n)\} \quad (5) \end{aligned}$$

There are also four types of confidence measures classified as grammatical and combined confidence measures [8].

N-gram Probability is a grammatical confidence measure. It is computed by re-scoring the output of recognizer (Viterbi phone sequence) conditioned by the n-gram language model probability of a history h . The length of h depends on used language model probabilities (2 for bigram, 3 for trigram).

$$nNG(q_k) = \frac{1}{D} \log\{P(q_k|h)\} \quad (6)$$

History contains previous decodings. Probability $P(q_k|h)$ is the probability of current decoding given history h and is calculated over language model. The confidence measure is obtained by normalizing the log value of this probability.

N-gram based Posterior Probability is a combined measure. It can be calculated by multiplying the posterior probability resulted from

Viterbi state sequence by the n-gram language model probabilities for current state sequence and a history h which includes previous decodings. The class prior probabilities are included in the calculations.

This confidence measure is a combination of Scaled Likelihood confidence measure and N-gram probability confidence measure.

Lattice Density is the average of the number of decoding hypothesis in the n -best decoding list for the interval selected for confidence measure, for a particular hypothesis. Lattice is formed by intervals represented by a frame. For each interval the number of competing hypotheses in the n -best list is counted and averaged over a duration.

LM-Jitter For this measure we have m different language model weighing factors. The n -best list obtained by the decoder is re-scored with the m language model weighing factor to obtain an m -best decoding list. This confidence measure, $LMJ(q_k)$, is obtained by counting the number of times that q_k remains at the same position in the m -best list.

To obtain the phone level confidence measures it is sufficient to take the interval chosen as phone. That means the beginning frame of the phone is n_s and the ending frame of the phone is n_e .

For word level confidence measures there are two possibilities. The first one is to use the same method like the measures for phone and the second one is to calculate phone-level confidence measures for all the phones of the word and then sum and normalize with the number of phones in the word. If the number of phones in the word is L , the Posterior Probability confidence measure can be calculated by;

$$nPP(w_j) = \frac{1}{L} \sum_{l=1}^L nPP(q_l)$$

III. MORPHOLOGY OF THE TURKISH LANGUAGE

The Language model is used to calculate a probability for word sequences. The use of language model with the acoustic model may be useful in searching for the best word sequence. Classical language models are based on words. Since The Turkish language is an agglutinative (inflected) language the number of words is very high and it is difficult to provide a language model.

Turkish has an agglutinative morphology with productive inflectional and derivational suffixations. Since it is possible to produce new words with suffixes, the number of words is very high. According to [4], the number of distinct words in a corpus of 10 million

words is greater than 400 thousand. This work also revealed the data sparseness problem for such a large corpus. Such sparseness increases the number of parameters to be estimated for a word based language model.

In agglutinative languages, it is possible to add morphemes after another one. Each morpheme conveys some morphological information such as tense, case, agreement etc. This property of agglutinative languages results in a large number of words which have the same stem but different endings.

Since there are so many words in the dictionary, if we use a large vocabulary continuous speech recognition system for Turkish language, there will be a large number of OOV words which are not modeled. An example for word production in Turkish [6];

OSMANLILAŞTIRAMAYABİLECEKLER İ
EKLERİMİZDENMİŞSİNİZCESİNE

This word can be broken down into morphemes as follows:

OSMAN+LI+LAŞ+TIR+AMA+
+YABİL+ECEK+LER+İMİZ+DEN+
+MİŞ+SİNİZ+CESİNE

The meaning of this word is “as if you were of those whom we might consider not converting into an Ottoman”. There is a study for the description of Turkish morphology by finite state approach [6]. In this work morphological rules of Turkish language are re-defined with finite state approach. The rules are compiled as finite state acceptors.

The phonemes in the Turkish Language each letter corresponds to a particular phoneme;

Vowels..... : a e ı i o ö u ü

Consonants: b c ç d f g ğ h j k l m n p r s ş t v y z

Language modeling for agglutinative languages need to be different than modeling languages like English. Such languages also have inflections but not to the same degree as an agglutinative language. The technique “Part-Of-Speech” (POS) which tries to build trigram models for word classes, is not good for agglutinative languages. Because there is a very large number of words which have the same root word. POS tries to define tag sets for endings and uses them in trigram language model.

One approach for modeling agglutinative languages is proposed by [3]. In this work the roots and the endings of words are considered as language model entries. The procedure is;

1. Identify all possible endings for a Language.
2. By using a particular vocabulary, extract the endings (Either by using a dictionary in which

- endings and stems are already defined or by processing the text to find endings and stems).
3. Take a sufficiently large text and generate a text composed from stems and endings separated by a space.
 4. Update the vocabulary used in step 2. by including the stems and endings found in step 3.
 5. For each stem calculate a set of probability for the endings.
 6. Generate a LM for any combination of stems and endings.

Another method for language modeling of agglutinative languages is proposed by [4]. This method is based on morphological structure of Turkish language. The aim is to model the distribution of morphological parses given the words and to seek a variable T , that maximizes $P(W|T)$:

$$\operatorname{argmax}_T P(T|W) = \operatorname{argmax}_T \frac{P(T)P(W|T)}{P(W)} \quad (7)$$

$$= \operatorname{argmax}_T P(T)P(W|T) \quad (8)$$

Since $P(W)$ is constant it can be ignored. T is the sequence of morphological parses for a word. T includes the root form of word and all morphosyntactic features to determine the word so the probability $P(W|T) = 1$ and the equation above is;

$$\operatorname{argmax}_T P(T|W) = \operatorname{argmax}_T P(T) \quad (9)$$

and as trigram tag model $P(T)$ can be defined as;

$$P(T) = \prod_{i=1}^n P(t_i|t_{i-2}, t_{i-1}) \quad (10)$$

where $P(t_1|t_{-1}, t_0) = P(t_1)$, $P(t_2|t_0, t_1) = P(t_2|t_1)$. Each t is a morphological structure and can contain one root word and a number of inflectional groups (IG). There are also some other simplifications to alleviate the data sparseness problem, like; root word depends only on the other root words.

The language models proposed in this work includes some trigram models for root words and inflectional groups.

IV. DATABASE PREPARATIONS

Turkish is not studied very much for speech recognition. The first explanation of this fact is probably due to a lack of databases in that particular language. For this reason the first step for Turkish speech recognition is database preparation. The texts to record the database was selected from the television news and some journal articles. Selection criteria was to cover more subject and to create a phonetically balanced corpus.

A. Text Selection

There was two different types of data; isolated words and continuous speech. For isolated words data the 100 most frequently used words are selected [9]. For continuous speech 215 utterances are selected from 50 different subjects. The isolated words was for labeling purposes. Isolated words are segmented and labeled by hand and this segmentation will be used for the segmentation and labeling of continuous speech.

B. Recording

For recording 20 native speakers were selected, 10 male and 10 female. Speakers were asked to read the prepared texts. Normal recording time for one speaker was 25 minutes, 22 minutes for continuous speech and 3 minutes for isolated words. If there is an error in reading, speakers were asked repeat the sentence. First they were asked to read continuous speech, then isolated words by stopping after each word.

Speech was recorded in a quiet room for each speaker. Recording materials were a portable Sony DAT-recorder TDC-8 and a close speaking Sennheiser microphone MD-441-U. Speech was digitally recorded at 32kHz sampling rate in stereo quality. It was transferred to computer with a Zefiro digital sound card and copied to CD-ROM environments for further processing.

C. Data Processing

The recorded speech was down-sampled to 16kHz with 16 bit resolution in mono quality. The utterances are separated by a wave editor, the erroneous utterances are eliminated. The isolated words are grouped by 10 and erroneous words are deleted. As a result there was 215 separated utterance for each of 20 speakers and 100 isolated words grouped by 10.

The isolated words for 10 of the speakers (5 male, 5 female) were segmented and phonetically labeled. The phoneme labels are selected as the letters used in written Turkish texts. (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). Those letters which differ from the letters of English alphabet are changed to the most similar English letter but they are presented in uppercase. This is done to simplify the processing of texts. 8 new phonemes were added to label the phonemes with a stop (b1, c1, C1, d1, g1, k1, p1, t1). The symbol for silence was chosen as #. The total number of phonemes used for labeling was 38 including the silence. The original letters converted to uppercase letters are: ç : C, ğ : G, ı : I, ö : O, ş : S, ü : U.

The program **snorri**¹ was used for phonetic labeling of data. The labeling of remaining data will be done

¹<http://www.loria.fr/~laprie>

automatically by using Speech Training and Recognition Unified Tool (STRUT)² developed at Laboratory of Circuit Theory and Signal Processing (TCTS) of Faculté Polytechnique de Mons [2].

D. Statistics

utterances (continuous speech) : 215,
 # isolated words : 100,
 # words in continuous speech : 2160,
 # different words in continuous speech : 1564,
 # different words in all recording : 1618,
 # speakers : 20 (10 male, 10 female)

V. TRAINING AND RECOGNITION RESULTS

STRUT software is used for speech analysis, acoustic model training and speech recognition purposes. There are independent programs for each step of training and recognition processes. In this context, STRUT programs are used to make segmentation and phonetic labeling of speech data given some labeled data. The resulting labeled data can be used to label new data.

The recognition performance of a system can be calculated by counting the correctly recognized words.

TABLE I
RESULTS FOR ISOLATED WORD RECOGNITION

	MLP Training Rate	Word Error Rate
Training Data	87.9%	0.6%
Training Data (aligned MLP)	87.9%	1.0%
Test Data	87.9%	6.6%
Test Data (aligned MLP)	87.9%	6.7%

The Turkish database described in section IV was used for experiments. The first results are obtained by training an MLP with phonetically labeled isolated words from 10 speakers. There were 100 words for each speakers. Then the isolated words speech from 10 other speakers was used to test the speech recognition system. For each data set, first, the MLP trained with hand-segmented data, then, the MLP trained with aligned data was used. The test data will be used for some more training of the system. The resulting MLP will be used to label the continuous speech. The MLP training rate obtained with the isolated word speech remains in the acceptable region. Confidence measures for incorrect words will be tested. Recognition accuracy for continuous speech should be good after some more training and introducing a well defined language model.

²<http://tcts.fpms.ac.be/asr/strut.html>

VI. CONCLUSIONS

Use of Hybrid HMM/MLP speech recognition system makes confidence measurements easier since they use discriminative training which use Maximum A Posteriori (MAP) criterion. Another advantage is that they allow easy normalization over posterior probabilities. They do not require the use of an additional model.

Confidence measures are important because there is a need for verification of output. It is possible to use confidence measurements as performance measure of a speech recognition system and it is possible to increase the performance of a speech recognition system with well defined confidence measures.

Since the languages are very different from each other it is important to use language specific properties in a speech recognition system. This properties can be used to select a training database, to create a language model which use the specific morphological structure of language, to use language specific properties in confidence measures.

Confidence measures are mainly used for testing hypotheses on recognizer outputs. These hypotheses can be, correctness of a given decoding hypothesis for an utterance (utterance verification), selection of correct word spoken in an acoustic signal including non-speech parts (keyword spotting) and detection of OOV words (OOV word spotting). In the case of hypothesis testing, confidence measures can be used to accept or reject a hypothesis according to a defined critical (threshold) value. In addition, it is possible to use confidence measures to filter a noisy signal or to search for best decoding [8].

REFERENCES

- [1] G. Bernardis and Bourlard H., Improving posterior based confidence measures in hybrid hmm/ann speech recognition systems. Technical report, IDIAP, 1998.
- [2] O. Deroo. *Modèles dépendants du contexte et méthodes de fusion de données appliqués à la reconnaissance de la parole par modèles hybrides HMM/MLP*. PhD thesis, Laboratory of Signal Processing and Circuit Theory, Faculté Polytechnique de Mons, 1998.
- [3] Kanevsky et. al. Statistical language model for inflected languages, 1998. US patent no: 5,835,888.
- [4] D. Hakkani-Tur, K. Oflazer, and G. Tur. Statistical morphological disambiguation for agglutinative languages. Technical report, Bilkent University, 2000.
- [5] S. Kamppari and T. Hazen. Word and phone level acoustic confidence scoring. In *ICASSP00*, pages 1799–1802, 2000.
- [6] K. Oflazer. Two-level description of turkish morphology. *Literary and Linguistic Computing*, 9(2):137–148, 1994.
- [7] G. Williams. A study of the use and evaluation of confidence measures in automatic speech recognition. Technical report, Department of Computer Science, University of Sheffield, 1998.
- [8] G. Williams. *Knowing What You Don't Know: Roles for Confidence Measures in Automatic Speech Recognition*. PhD

thesis, Department of Computer Science, University of Sheffield, 1999.

- [9] C. Yilmaz. A large vocabulary speech recognition system for turkish. MS thesis, Bilkent University, 1999.