Confidence Measures For Turkish Call Center Conversations

Ali Haznedaroglu 1,2, Levent M. Arslan 1,2
1 Electrical and Electronics Engineering Department, Bogazici University, Istanbul, Turkey
2 Sestek Inc., Istanbul, Turkey

{ali.haznedaroglu, arslanle}@boun.edu.tr

Abstract

Automatic speech recognition accuracies of call center conversations are still below intended levels due to harsh conditions such as channel distortions, external noises, co-articulated speech, etc. Agglutinative and free word order nature of Turkish degrades the recognition performances further; therefore the usage of confidence measures (CMs) is inevitable to retrieve correct information from the calls. In this paper, two conversational CMs, namely speech overlap ratio and opposite party energy level, are proposed, and tested together with single-channel confidence measures on Turkish stereo call center recordings. Experimental results show that conversational CMs improve the rating accuracies of the utterances with respect to their recognition rates.

Index Terms: confidence measures, large vocabulary continuous speech recognition, information retrieval

1. Introduction

Call centers have huge impact on customer satisfaction for companies, since they act as the primary intermediary between the company and its customers for product support or information inquiries. Tracking the calls of a call center may give valuable information to the company, such as performance of its agents, satisfaction of its customers and analysis of the campaigns and competitors. As manual call tracking is an expensive process both time and resource wise, automatic retrieval of call information is an attractive choice.

In this work, we present a Turkish large vocabulary continuous speech recognition (LVCSR) system that acts on real call center data, and we implement confidence measures (CMs) in order to increase the reliability of the statistics inferred from the system.

Large vocabulary continuous speech recognition systems are commonly used for speech mining in call center applications as they can give rich speech to text outputs that can be used for many different information retrieval purposes. However, speech recognition performance of such systems degrade due to some harsh conditions such as telephone channel distortions, external noises on both customer and agent channels, huge voice and speaking style variability of the customers, disfluent co-articulated speech, and high perplexity of largely varying customer request content [1, 2]. The degradation is even worse for Turkish systems because of the language's agglutinative and free word order nature. In such situations, confidence measures (CMs) can be used to evaluate the reliability of the recognition results, and take corrective action for the words or utterances that are likely to be erroneous. Speech recognition systems can hugely benefit from confidence measures as these measures play a crucial role in usefulness of such systems in practical, real world scenarios.

Previous efforts on telephony speech based CMs mainly deal with assigning confidence scores for individual words, such as certain dialog slots, or important keywords in automated spoken dialog systems. In [3] normalized word likelihoods at the end of Viterbi decoding, number of alternative word hypothesis above the threshold when the word ends, number of phonemes in the word, and the word duration are used as word-based CMs on the Switchboard Corpus. Acoustic stability is also evaluated as a CM on the same corpus in [4]. Language model scores, back-off behaviors and phonetic length of the recognized words are chosen as the confidence measures, and their combinations using both multi-layer perceptron and a statistical decision tree are also investigated in a telephone-based spoken language system that accesses airline information in [5]. In [6], confidence of a hypothesized word is estimated directly from its posterior probability in human-to-machine dialogs over the telephone. Other than features obtained from the decoding stage, semantic information is also used in telephony applications as CMs in [7, 8, 9].

In this paper, we investigate CMs on the LVCSR outputs of real Turkish call center conversations. Our aim is to automatically filter out call segments that have poor recognition accuracies in order to retrieve correct statistics from the calls. In addition to some conventional single-channel CMs, we propose new conversational measures which incorporate information from both channels of the stereo telephone conversations, namely speech overlap ratio and opposite channel energy level. In order to combine the features and rate the call segments, we use support vector machines (SVMs).

In the following two sections, our database and LVCSR system is presented. Then, we explain the confidence measures used and their combinations in the forth section. Finally we present our experimental results, and discuss them.

2. Database

Our database consists of approximately 200 hours of real human to human call center conversations in Turkish, collected from different domains such as banking, insurance and telecommunications. The calls are initially recorded in stereo, 8 kHz, 8-bit, A-law format. They are separated into their agent-customer channels, and converted into 16-bit PCM format for training and test purposes. All of the calls are manually transcribed for agent and customer channels separately. Nearly 12 hours of insurance conversation data is left as the test set, of which 9 hours is used in SVM training, and the rest is used in tests.

3. LVCSR system

Sphinx-3 LVCSR engine [10] is used in our acoustic model training and decoding. Feature vectors consist of 12 Mel-frequency cepstral coefficients and energy, together with their first and second order derivatives. Cepstral mean normalization is applied to the feature sets in order to minimize the transmission channel differences. Context-dependent triphone models with 6000 tied-states using 12...
Gaussian mixtures are then trained for a Turkish set of 29 letters. Since the orthographical and phonetic transcriptions significantly match in Turkish, we decided to use letters instead of phonemes as the modeling units. Silence and additional filler models, such as DTMF, lip and background noises, are also trained. Recordings of both agent and customer channels are used together in acoustic model training.

Language model training is done using the CMU SLM Toolkit [11]. Different language models are trained for customer and agent channels of the target call center because of the lexicon and jargon differences amongst them. Both word based bigram models, and subword based trigram models [12] are trained for our test purposes. Subword models are constructed from stems and endings of words [13], which are obtained using Morfessor, which is an unsupervised algorithm [14]. Subword models are used for acoustic stability CM tests in [15], but they are not included in this work. Unigram and bigram counts for the word based bigram language models are given in Table 1.

<table>
<thead>
<tr>
<th></th>
<th># Unigram</th>
<th># Bigram</th>
</tr>
</thead>
<tbody>
<tr>
<td>Agent</td>
<td>12512</td>
<td>71691</td>
</tr>
<tr>
<td>Customer</td>
<td>13346</td>
<td>68048</td>
</tr>
</tbody>
</table>

Table 1: N-gram counts of the language models

4. Confidence measures

CMs that are used in this work are categorized into two groups. First group is the single-channel conventional measures, which are calculated using single channel of the call center conversations, i.e. only the agent channel or the customer channel. The other group, which we propose in this study, consists of conversational two-channel CMs, which also take into account the opposite channel of the conversations. All the confidence measures are calculated at the utterance level, which are automatically segmented from each call using energy-based voice activity detection (VAD) module. Since a separate CM is calculated for each utterance, noisy or problematic parts of a conversation can be eliminated while the remaining parts are used for information retrieval purposes. In the next two subsections, we will explain these CMs.

4.1. Single-channel CMs

Single-channel CMs are calculated using the decoder output and prosodic features of the utterances.

- **AM score**: Acoustic model score obtained from Viterbi decoding that is normalized by the number of frames.
- **LM score**: Language model score obtained from Viterbi decoding that is normalized by word count [5].
- **Posterior probability**: Total hypothesis score that is normalized by the acoustic data observation probability [16]. To determine the acoustic observation probability, we run a parallel Turkish all-phone recognizer in which all monophone, and filler HMMs are fully connected together without any restrictions in the grammar network.
- **Prosodic features**: These features include median and standard deviation of pitch and average energy [17].

4.2. Conversational two-channel CMs

In examining call center conversations, we observe that some speech features of the conversation’s other party can affect speech comprehensibility of the current party, and may cause degradation in speech recognition performance. For this reason we propose the following confidence measures, which are called conversational CMs.

4.2.1. Speech overlap ratio

Simultaneous talk on both channels of the conversation can affect the recognizability of the speech on each channel. When the other party of the call interrupts their speech, people tend to hesitate which degrades their speech quality.

Also, when the other party of the conversation is simultaneously talking, people involuntarily tend to increase the intensity of their voice to enhance their audibility, known as the Lombard reflex. Lombard reflex not only increases the energy of the speech, but it also increases the phonetic fundamental frequencies, and shifts the formants. Energy is shifted from low frequencies to middle or higher frequencies resulting in spectral tilt. Duration of the vowel sounds and content words are also increased. Automatic speech recognition systems are sensitive to the Lombard reflex, due to its resulting increase in the auditory SNR of the speaker’s spoken words [18].

In order to calculate the speech overlap ratio for an utterance, the recognition outputs of both channels of the conversation are time-aligned first. Then, corresponding overlapped portions of speech utterances of both channels are determined and speech overlap ratio of an utterance is found by dividing this overlap duration by the total duration of the utterance. For example, speech overlap ratio of the customer utterance in Figure 1 is 0.2.

![Speech overlap ratio example.](image)

4.2.2. Energy level of the opposite party

Energy level changes on opposite parties of the conversation may also affect intelligibility, and hence the automatic recognition output. We observe that when the other channel’s voice level is low in telephone conversations, speakers tend to speak louder, which may decrease the recognition accuracy. For this reason, we also include the average energy of opposite’s party speech as a confidence measure.

4.3. Combination of CMs

Various fusion techniques were implemented in the literature for confidence measures. In this work, we use support vector machines to combine the single-channel and conversational two-channel CMs as in [19]. SVMs constitute supervised learning methods to analyze and classify data. In this method, training vectors are mapped into a higher dimensional space using kernel functions, and SVM finds a separating hyper-plane with the maximal margin.
in this higher dimensional space. LIBSVM library [20] is used in our simulations.

5. Experiments

In our experiments, the aim is to automatically rate the test utterances with respect to their word error rates (WERs) using the confidence features explained in the previous section. Correct detection of highly misrecognized utterances will enable us to filter them out, in order to retrieve more reliable statistics from the calls.

5.1. LVCSR baseline

Twelve hours of stereo Turkish call center conversations are excluded from acoustic and language model trainings. These calls are separated into their agent and customer channels, and each channel recording is automatically segmented into utterances using an energy-based VAD module. Then each utterance is automatically transcribed using the Sphinx-3 decoder [10], and confidence features explained in the previous section are extracted.

Nine hours of this dataset is used in the SVM training for different CM combinations, and 3 hours are set aside as the SVM test set. Speech recognition accuracies of the test sets for the agent and customer sides are given in Table 2.

<table>
<thead>
<tr>
<th></th>
<th>Agent WER %</th>
<th>Customer WER %</th>
</tr>
</thead>
<tbody>
<tr>
<td>Single-channel CMs</td>
<td>32.74</td>
<td>53.50</td>
</tr>
</tbody>
</table>

Table 2: SR results of the test set.

High error rates on both channels are mainly caused by the OOV words due to highly varying speech content, speaker or speaking style variability, and low-quality recordings from the telephone channels. Agent channels have better recognition accuracies since the agents are trained to talk fluently, and their vocabulary is more constrained. Also, in general there is less background noise in agent channel recordings when compared to customer channels. These high WERs make the use of CMs inevitable in such practical applications.

5.2. SVM training

All the utterances in the SVM training set are first sorted with respect to their recognition error rates in descending order. Then, this list is divided into three parts having equal train set coverage in order to balance the data between good and bad recognition results. Utterances in the first third and the last third of this list are labeled as “reject” and “accept” respectively, and they are included in the SVM training phase. In this stage, we do not include the middle portion in order to train more discriminative models for the CMs by using only the utterances with high and low WERs.

SVM training is done using the LIBSVM library [20]. Non-linear radial basis function (RBF) kernel is used in the trainings and best kernel parameters are found using cross validation. Then, SVM models are trained using these parameters.

SVM training and tests are done for 4 different CM combinations. Our baseline set consists of the single-channel CMs presented in subsection 4.1. Then each of the conversational two-channel CMs in subsection 4.2 is individually added to this baseline set, constructing the other two sets. Our last set includes all of the single-channel and conversational two-channel CMs.

5.3. SVM tests

SVM test gives a probability score for each test utterance, where probability of 1 indicates that the utterance has high confidence of having good recognition accuracy, and probability of 0 indicates otherwise. All of the agent and customer test utterances are sorted in descending order with respect to these probability scores separately. Then these sorted lists are divided into three equal portions, having the test set coverage of 33.3%. In order to rate a CM set more successful, we require that the upper portion has a lower WER, while the lower portion has higher WER. Results for these portions and different CM sets are given in Table 3, and also in Figures 2 and 3. Theoretical limits indicate the real WERs of the test set when it is divided into three equal portions with respect to the recognition accuracies.

<table>
<thead>
<tr>
<th></th>
<th>Agent WER</th>
<th>Customer WER</th>
</tr>
</thead>
<tbody>
<tr>
<td>Single-channel CMs</td>
<td>54.44</td>
<td>78.50</td>
</tr>
<tr>
<td>Single-channel CMs + opposite energy</td>
<td>54.52</td>
<td>78.55</td>
</tr>
<tr>
<td>Single-channel CMs + speech overlap</td>
<td>55.51</td>
<td>78.59</td>
</tr>
<tr>
<td>Single-channel + Conversational CMs</td>
<td>55.71</td>
<td>79.08</td>
</tr>
<tr>
<td>Theoretical limit</td>
<td>63.36</td>
<td>89.15</td>
</tr>
</tbody>
</table>

Table 3: WERs of different CM sets.

6. Discussion

From the test results, we see that best utterance rating is obtained when the conversational two-channel CMs are added to the single-channel CMs. Within the conversational CMs, speech overlap ratio gives slightly better results than the opposite energy level. We also observe that agent-channel rating accuracy improves more than the customer-channel rating accuracy when the conversational CMs are used. The reason for this can be that, agents are trained to be more polite and permissive, and they hesitate more when the customer interrupts their speech.

Although using conversational two-channel CMs together with the single-channel CMs increase the rating accuracy of utterances for both high and low recognition results, the performance is still well below the theoretical limits.
7. Conclusions

In this paper, we build a Turkish LVCSR system that acts on real call-center conversations and we investigate the performance of confidence measures on the recognition outputs. Confidence measures are required in such practical applications where the WERs are high. Main contribution of this work is the conversational two-channel confidence measures, which also take into account acoustic features of the opposite party’s speech in a call center dialog.

Conversational CMs proposed here are the speech overlap ratio, which takes into account the relative duration of simultaneous speech on both channels, and the energy level of the opposite party. These CMs are mainly based on our observations on large amount of call center conversations. Conversational measures are combined together with the conventional single-channel measures using SVMs, and their accuracies in rating the utterances with respect to the recognition accuracies are tested.

WERs of the LVCSR system are found to be 32.74% and 53.50% for the agent and customer test utterances, respectively. SVM training and tests are done for 4 different CM combinations. Our baseline set consists of the single-channel CMs only. Then, combinations of the conversational CM combinations. Our baseline set consists of the single-channel CMs + opposite side energy. Single-channel CMs + overlap ratio. Single-channel CMs + conversational CMs.

Figure 3: WERs of customer channel for different CM sets.

Although conversational CMs used in this paper are shown to improve the rating accuracies, the performance is still well below the theoretical limits. In our future work, we plan to incorporate other high level confidence features in addition to the ones presented here. One of the features that will be investigated is the semantic information of the opposite party in the conversation. We also plan to integrate an emotion recognition engine and use its output as a confidence feature of the utterance. From our observations on call-center conversations, angry customer’s speech becomes incomprehensible due to the changes in its energy and pitch characteristics.

8. Acknowledgements

This work is partially funded by TUBITAK within a TEYDEB project number 3070164.

9. References