Automated Transcription of Conversational Call Center Speech – with Respect to Non-verbal Acoustic Events

Gellért Sárosi¹, Balázs Tarján¹, Tibor Fegyó¹,², and Péter Mihajlik¹,³
¹Department of Telecommunication and Media Informatics, Budapest University of Technology and Economics, Hungary; ²Aitia International Inc.; ³THINKTech Research Center Nonprofit LLC

Abstract. This paper summarizes our recent efforts made to transcribe real-life Call Center conversations automatically with respect to non-verbal acoustic events, as well. Future Call Centers – as cognitive infocom systems – must respond automatically not only for well formed utterances but also for spontaneous and non-word speaker manifestations and must be robust against sudden noises. Conversational telephony speech transcription itself is a big challenge, primarily we address this issue on real-life (Bank and Insurance) tasks. In addition, we introduce several non-word acoustic modeling approaches and their integration to LVCSR (Large Vocabulary Continuous Speech Recognition). In the experiments, one and two channel (client and agent speech merged into one or left in two separate audio stream) transcription results, cross-task results and the handling of transcription data insufficiency are investigated – in parallel with the non-verbal acoustic event modeling. On the agent side less than 15% word error rate could be achieved and the best error rate reduction is 20% (relative) due to the inclusion of various written corpora and due to acoustic event handling.

1 Introduction

Call Centers produce a vast amount of speech that need to be stored but typically not processed further. Client-agent conversations are invaluable for the companies as they convey direct information from the "first line" – from the interface between the company and their clients. Obviously, these speech data contain direct client feedback, opinions, complaints and sentiments. Furthermore, it is essential that the agent – as the "face" of the company – be closely monitored and controlled. There are several difficulties in processing Call Center speech: the first priority, the speech-to-text transcription itself is a big challenge due to the spontaneous, conversational style, to the diverse speaking style and rate, and of course, to the background and transmission noises. Second, the speech contains far more information beyond simple text content, and even the borderline is not straightforward to be drawn between "normal speech" and "meaningful but not phonetic, speech-like content" and other "speaker noises" that may have importance in the communication.

In this paper we make an attempt to improve the Large Vocabulary Continuous Speech Recognition (LVCSR) of Call Center speech and to integrate the non-verbal and non-speech acoustic events to the recognition process. To the best of our knowledge, in this genre, this is the first scientific paper discussing Hungarian Call Center speech recognition. Besides, one of the lowest Hungarian LVCSR word error rates are achieved on real-life (Bank and Insurance Call Center) tasks.

Our current aim was to develop a real-time speech transcription system for Call Center task that is able to roughly transcribe the conversations. The targeted application is monitoring and content analysis/text mining of the recorded speech. This application have to process hundreds of hours of telephone speech data generated in a call center (Figure 1), then text analytics are applied on the automatic transcriptions, and multiple reports are generated to help improving client retention, agent efficiency, marketing, etc. In this paper, we focus on developing the speech recognition module, to generate richer and more accurate transcriptions. For such an application the recognition of non-word acoustic events can be highly beneficial, as, for example, some of these events
2 Related work

Automatic transcription of Call Center speech and detection of non-verbal vocalizations represent two separate research fields. As there are only few works investigating these fields jointly, the studies cited in this section are grouped according to their topics.

2.1 Recognition of Call Center speech

Call centers are particularly important contact points between the companies and their clients, as they are the clients’ first interface that has a high effect on company’s reputation. Many efforts have been made in development of automated call center transcriptions and quality monitoring systems, thus only those are cited here that relates the most to our work. In [1] the potentials of exploiting unsupervised adaptation techniques for automated call center speech transcription were investigated. The initial tests were performed using language and acoustic models of Switchboard system which were first replaced with models trained in unsupervised manner and then with task-specific models. The initial WER of the system was 53.3% for caller and 54.2% for agent data, which fell to 43.5% and 27% respectively.

Zweig and his colleagues describe an automated system for assigning quality scores to recorded call center conversations [2]. They used two types of quality estimations. First simple pattern matching on the ASR transcript to answer a set of standard quality control questions and then maximum entropy ranking, where their goal was to estimate the probability of a call being bad based on features extracted from the automatic transcription. Both approaches were evaluated on a 3 hours test set from IBM’s North American call centers. Maximum entropy ranking was outperformed the question answering technique in terms bad call retrieval accuracy, however their combination was the most successful.

As a part of the CALLSURF project 20 hours of call center conversational data were transcribed in high quality and a further 150 hours with less strict transcription rules [3]. The baseline system used LIMSI conversational telephone speech (CTS) recognition system to transcribe 10 hours of call center speech and 47% WER was reported. CTS system was refined by adapting the acoustic models using the fine CALLSURF transcripts, and using all the available (fast and fine) transcripts for language modeling. The word error rate was reduced to about 35%.

The above cited papers describe various approaches to create high quality automated call center transcription systems, however there is place for further improvement. There can be significant acoustic and lexical differences between agent and client speech in call centers, which is not exploited by any of the systems. This study hence makes an attempt for adapting the recognition models on agent and client side basis. Although some results can be found in [1] for cross evaluation of call center recognition tasks, in our paper it is performed on a much less resourced task and additional textual training data are adapted to the final system.

2.2 Recognition of non-verbal vocalizations

In an early work by Schultz and Rogina [4], non-verbal sounds are modeled similarly to normal phones in order to improve the robustness of speech recognition, however detection rates for non-verbs are not reported. Some later studies ([5], [6], [7]) focus on the detection of specific non-verbs (laughter, hesitation). [5] uses F0 and spectral envelope as features for hesitation detection and achieves 85% recall and 91.5% precision. [6] analyzes different types of laughter and suggests a solution for classifying them. In contrast, [7] steps a bit further and introduces a neural network based classifier for discriminating laughter from speech. The best results (Equal-Error Rate of 7.9%) were achieved by using Mel-Frequency Cepstral Coefficient (MFCC) as features.

A comprehensive study by Schuller and his colleagues...
[8] makes an attempt to recognize various non-verbal vocalizations (breathing, consent, coughing, hesitation, laughter and other human noise), whereas comparing the performance of three types of classifiers: Hidden Markov Models (HMM), Hidden Conditional Random Fields (HCRF), and Support Vector Machines (SVM). In the absence of prior experience the topology and parameters of HMMs are also optimized in this work. It was found that HMMs with linear topology, high number of states (above 5) and 8 Gaussian mixture components can outperform HCRF and SVM based approaches. Although perceptual linear predictive (PLP) features performed slightly better than MFCCs, the difference was very small.

The approach described in [8] is further refined in [9] and [10]. In [9] Non-negative Matrix Factorization (NMF) technique is applied to achieve new features for discrimination of speech, laughter, breathing, hesitation or non-verbal consent. NMF features by themselves provide less accurate classification than MFCCs, however combination of the two features sets resulted in a slight improvement. [10] investigates the potentials of using artificial neural networks especially Bidirectional Long Short-Term Memory Recurrent Neural Networks (BLSTM-RNNs) instead of HMMs in the classification task of non-verbal vocalizations. BLSTM-RNNs reduced average error rate from 8.7% to 6.3%. BLSTM significantly outperformed HMMs for non-human noises, whereas for laughter HMMs were found to be more accurate.

All the above approaches (except for [4]) operate outside the Automatic Speech Recognition (ASR) framework and hence be used for a two-stage decoding process. Our aim was to build a real-time, one-stage speech transcription system, thus non-verbal event detection had to be integrated into the ASR system. For that reason non-verbals are modeled with HMMs here, which is suboptimal for some non-verbals (especially for non-human noises) compared to neural networks, however fits well with ASR framework. Although some of the cited studies suggest other feature sets, we kept at MFCCs, since it has a balanced performance among speech and non-verbals.

3 Call Center Databases

We collected and transcribed various amounts of Hungarian audio data from companies’ Call Centers. Table 1 shows the basic properties of three audio databases. A dataset was received from an Insurance Call Center referred as Insur, and a database was came from Banking Call Center, referred as Bank. Another database was received from the same Company (but from different branch) as the source of Insur. The development and test methodology on this dataset did not allow scientific experiments, however we used this training data – referred as the Supplementary – in combination with the other two. These databases contain sensitive information about clients, therefore both companies requested complete secrecy, meaning that it is not allowed to give any audio or textual data to a third party.

The Insur corpus was recorded in two channel format, thus there is no overlap between the speech of the agent and the client since they are recorded in two separate streams – only, occasionally acoustic crosstalk caused some disturbance. This corpus contains outbound calls – surveying clients, making new contracts, or expanding existing ones – that were previously agreed by clients. During manual transcriptions, precise segmentation was conducted and speech-like non-verbal acoustic events were labeled as displayed in Table 2 using Transcriber1. Because of the nature of these calls, the numbers of inter-word emotional manifestations and communicative expressions were frequent. Some examples situations are:

- The client hesitates while thinking about personal data during identity verification
- Hesitation, agreement or denial expressions, while the client thinks or decides about offered services or solutions to the problem
- Spontaneous expressions for various degrees of discontent

Inter-word and background non-speech content was only annotated during the early phase of the transcription work, but later we changed it to make it more economic. As it is reflected in Table 2 only the non-verbal events were annotated, and longer segments containing only noises were labeled as nonspeech.

The Bank database consists of both incoming and outgoing calls in single channel format. The call types has a wide profile including money transfer requests, contract administration, marketing new services. For the text transcription, we had to follow the economic method of the Insur corpus, but we received other text sources – emails,

---

1http://trans.sourceforge.net/
Table 2: Transcribed non-verbal expressions and non-speech events

<table>
<thead>
<tr>
<th></th>
<th>Non-verbal</th>
<th>Non-speech</th>
</tr>
</thead>
<tbody>
<tr>
<td>Insur early</td>
<td>hesitation (ee, aa, mm) consensus (aha, mhm)</td>
<td>breath, laugh, cough, sneeze lipsmack, ring, keyboard, noise, music speech, crosstalk</td>
</tr>
<tr>
<td>Insur final</td>
<td>hesitation (ee, aa, mm) consensus (aha, mhm)</td>
<td>Not transcribed</td>
</tr>
<tr>
<td>Bank</td>
<td>hesitation (ee, aa, mm) consensus (aha, mhm)</td>
<td>Not transcribed</td>
</tr>
<tr>
<td>Supplementary</td>
<td>hesitation (ee, aa, mm) consensus (aha, mhm)</td>
<td>breath, laugh, cough, sneeze lipsmack, ring, keyboard, noise, music speech</td>
</tr>
</tbody>
</table>

agent manuals – from the company to help model personalization. See details in Section 5.

The Supplementary corpus consists solely of single channel recordings (the speech of the agent and the client are mixed into one stream) of inbound calls – client inquiries, complaints, requests. These were more rich in non-verbal and non-speech acoustic events, therefore it was transcribed entirely using the less economic but more detailed method of the early Insur transcription work.

Each database was stored in ADPCM format [11], but they were all converted to 8kHz sample rate, 16 bit linear encoding WAV files. The single channel data was kept as single channel, and the two channel data was separated to two separate single channel files.

4 Acoustic modeling

4.1 Acoustic front-end

Standard MFCC feature extraction with BEQ (Blind Equalization [12]) was applied on all audio data, including first and second derivatives and energy, resulting in a total of 39 dimensional vectors.

4.2 Training methodology

The audio data is segmented according to their transcriptions, and the segments with content unfit for training are excluded, then a standard "flat start" training – a ML (Maximum Likelihood) approach [13] – is performed. First a monophone 1 GMM model was trained on a subset of the corpus, then a forced alignment was performed to sort out the segments not fitting the initial model. After monophone training on the full but filtered database, the model was triphonized and three-state left-to-right HMM (Hidden Markov Model) was taught for each context-dependent phone model. At the end, across-word shared state triphone models were produced using phonetic decision trees, and the mixture number was incremented to an optimal value between 10 and 15 mixtures, with the state numbers around 5000-6000. For the whole training process various tools of the HTK toolkit [13] are used.

4.3 Baseline noise modeling

In the basic modeling approach, all speech-like, non-phonetic acoustic events were simply ignored, but the non-speech acoustic events were kept. All speaker related noises (like "breath", "cough", etc.), and the background "speech" event were grouped together into one speaker noise label to avoid under-training. The "ring" event is a rarely changing, special element of the call center context, therefore it remained separate from the more general "music" event. The "keyboard" event was grouped together with the "noise" event. This resulted in a total of four non-speech event categories intended for model training: speaker noises, music, ring, other noises.

The transcription rules changed in the early phase of the Insur work as shown in Table 2 to improve transcription efficiency, therefore the non-speech events were only annotated in a part of the training data. The Bank data transcription was entirely done in the most cost efficient way, thus it totally lacked noise annotations, moreover this was a significantly smaller database. To allow model training on labeled noise data, we had to add the Supplementary dataset to both the Insur and Bank databases. In addition, the Insur audio data was in two channel format, thus we could easily separate the agent and client voices – we also did this on the Supplementary dataset –, and we could easily separate the agent and client voices – we also did this on the Supplementary dataset –, and we could easily separate the agent and client voices – we also did this on the Supplementary dataset –, and were able to train separate acoustic models for the two channels, as well as a single model fit for both channels.

After the model training finished, the four non-speech GMMs were merged into the silence model, by simply moving the mixtures, thus increasing the silence model complexity. This way, output labels for noises do not appear during recognition, but the number of misrecognized words during noises should decrease. This approach was applied to train all Insur models and some Bank models.

4.4 Explicit modeling of non-verbal events and noises

Non-word expressions – hesitation, consensus, etc. – are a frequent part of the human communication, however, they were completely ignored during baseline acoustic training. To allow the recognition of these non-word events and to avoid confusions with regular word phonemes, an explicit non-verbal event modeling was required. All expressions were modeled with their unique
The non-speech content is also a frequent part of Call Center recordings. While the baseline approach can handle these noises in a simple, but efficient way, an explicit noise modeling approach was added during the last phase of the Bank acoustic modeling similarly to our approach in [14]. The <noise>, <speech> and <music> labels were all used to train their own distinct acoustic models. The speaker related noises – such as laugh, cough, breath, etc. – were grouped together for training a <spk_noise> model.

5 Language modeling

Building language models for morphologically rich languages (like Hungarian) is a challenging task due to data sparseness, high OOV rate and large lexicons [15]. These issues are usually handled by using subword language models instead of the standard word-based ones ([16], [17], [18]). However, in our former studies we found that the benefit from using subword models highly depends on the amount of available training data and the speech genre of the recognition task ([19], [20], [21]). According to these studies and a few preliminary measurements, we decided to use word-based models for call center speech, since the expected degree of improvement would not be high enough to compensate the increase in complexity. Consequently, by default for all the tasks word-based, trigram language models with Kneser-Ney smoothing [22] were built by using the SRI Language Modeling toolkit (SRILM) [23].

5.1 Training text corpora

As we already described in Section 3, three call center databases (Insur, Bank, Supplementary) were collected for training and testing each call center speech transcription system. These databases consist of manual transcriptions of call center conversations, which are in the case of Supplementary and Insur were also labeled by the type of speaker (agent/client). As manual transcriptions usually match well with the recognition task, they are widely used for training the language models of LVCSR systems. However, due to their high cost the amount of available transcriptions is usually not sufficient. In our research this problem rose for Bank task, where only 25 hours of speech were transcribed, hence we decided to utilize two additional text corpora, as well. The first additional corpus consists of e-mails from customers of the bank, whereas the second one is based on an agent training manual giving sample sentences serving as a guideline for call center agents. The size of each training text corpora can be found in Table 4.

Text normalization followed the standard approach when the baseline and noise merged acoustic models were applied. First the corpus was segmented into sentences, then non-verbal events (see Table 2) and special characters (e.g. comma, period, etc.) were removed and finally all the tokens (except for named-entities) were converted to lower case. In contrast, the non-verbal events were kept in the training text for those models that support event recognition.

5.2 Combination of call center language models

For training language models of Insur we also utilized the transcriptions of the Supplementary set in order to achieve a more robust model. The models esti-
mated on the Supplementary set and the Insur manual transcriptions were incorporated into a common model using linear interpolation [24] implemented in SRILM toolkit. Interpolation weight was optimized on a held-out Insur tuning set. As Insur is the only task where the agent and client speech were recorded on separate channels, this task was selected to investigate the potentials of adapting the language model for the two different speech types. Although the Supplementary database consist of one channel records, the origin of calls was labeled manually during the transcription process. The language model adaptation process for the agent/client channel of Insur included the following steps:

- Standalone agent and client language models are built for Insur task and Supplementary set
- Insur agent and Supplementary agent models are interpolated with a weight optimized on Insur agent/client tuning set
- Insur client and Supplementary client models are interpolated with a weight optimized on Insur agent/client tuning set
- The above two interpolated models are interpolated with a weight optimized on Insur agent/client tuning set

5.3 Integration of additional knowledge sources

Since only a small amount of manual transcriptions were available for the Bank call center task (see Table 4), we decided to focus on the integration of different knowledge sources to improve the baseline system. Although the usual topics in bank and insurance call centers considerably differ, the lack of in-domain training text prompted us to use every opportunity, thus we adapted both insurance models to the Bank baseline system by using linear interpolation. This model contains all manual transcriptions, but its parameters are optimized for the Bank recognition task. After all transcriptions were integrated into a common model, we were interested if this model could be further improved by using additional training data. Accordingly both the Bank e-mail corpus and the agent training manual were adapted one by one to the interpolated model of call center transcriptions. In the final language model each n-gram probability is calculated from the component models with the following weights: 0.7 for Bank transcriptions, 0.2 for Insur and Supplementary transcriptions, 0.05 for Bank e-mail corpus and 0.05 for Bank agent manuals. The effect of additional knowledge sources on recognition error can be followed up in Section 7.2.

6 Experimental setup

In all experiments the following techniques and environment was used beside the previously described acoustic and language modeling methods.

6.1 Pronunciation and context dependency models

To connect the phoneme HMM states of the acoustic model and the word labels of the language model, simple grapheme-to-phoneme rules [25] and an exception list were used to generate word to phoneme mappings. In the case of event models non-word utterances (\(<\text{ee}>, <\text{aa}>, <\text{mhm}>, \text{etc.}\) were mapped into their acoustic model labels. An optional silence model (similar to the ‘short pause’ (sp) model in [13]) were added to the end of each word like in Figure 2. In case of the baseline noise modeling approach, this ‘short pause’ only contained the \(<\text{sil}>\) and \(<\text{eps}>\) branches, while during the explicit non-speech modeling the various noise labels provided additional paths. The pronunciation models were further processed by applying triphone context expansion, as shown in Equation 1 below. This includes not only the inter-word dependencies but the cross-word context dependencies as well, taking the inter-word silences into consideration. Note that during triphone context expansion all non-word and non-speech event environments were considered as silence model environments.

6.2 WFST recognition networks

In the final step all the previously described knowledge sources are integrated into a triphone level WFST (Weighted Finite State Transducer) [27] recognition network. The strength of WFST framework lies in its ability to make the search space as compact as possible. The network construction and optimization scheme is detailed in Equation 1, where capital letters denote transducers, while the others are operators. The process commences with the composition and determinization of the language model (G) and the pronunciation model (L), then a sub-optimal minimization process is applied. The optional silences (sp) are replaced with inter-word silence model (S) transducer. Next, the context expansion is performed using the (C) transducer, then the network is minimized, factorized, and the weights are redistributed, resulting in a stochastic transducer suitable for a WFST decoder. All the operations were performed with “Mtool” WFST building toolkit.

\[
\text{wred}(\text{fact}(<\text{comp}(\text{C} \circ S \circ \text{comp}(\det(L \circ G)))))) \tag{1}
\]

The one-pass recognition tests were performed using the WFST decoder called VOXxerver [21] on a standard Linux PC with a Core2Duo processor at 3.16 GHz. Large
vocabulary speech recognizers are commonly characterized by their Word Error Rate (WER), however in the case of morphologically rich languages this metric shows a pessimistic picture of the speech recognition performance [26], hence Letter Error Rate (LER) is also given. In our LER calculation the white spaces between words were modeled by a dedicated letter. The Real Time Factor (RTF) of recognition tests was measured to be lower than 1, which means that our automatic call center transcription systems can process speech data faster than real-time.

7 Results and discussion

The aim of the Insur tests were to determine the benefit of two channel recognition and the effect of a VAD (Voice Activity Detector) front-end, and to examine the necessity of two channel modeling. The Bank tests focused on the low resource recognition task, and the effect of integrating additional knowledge sources into the existing models, and the addition of explicit event modeling. Table 5 summarizes the tested model sets and their techniques grouped by the recognition task.

Table 5: Non-verbal and non-speech event handling approaches in the recognition network

<table>
<thead>
<tr>
<th>Modeling of</th>
<th>Non-verbal expressions</th>
<th>Non-speech acoustic events</th>
</tr>
</thead>
<tbody>
<tr>
<td>Insur</td>
<td>Baseline 2-ch.</td>
<td>None</td>
</tr>
<tr>
<td></td>
<td>Baseline 1-ch.</td>
<td>None</td>
</tr>
<tr>
<td>Bank</td>
<td>Baseline</td>
<td>None</td>
</tr>
<tr>
<td></td>
<td>Event Models 1</td>
<td>Explicit</td>
</tr>
<tr>
<td></td>
<td>Event Models 2</td>
<td>Explicit</td>
</tr>
</tbody>
</table>

7.1 Two-channel recognition tests

Investigations were started right at the early phase of Insur data transcription, to see the 2-channel recognition performance in an under-resourced environment, in combination with a VAD (Voice Activity Detector). Similarly to the training corpus, the Insur outbound call center test data archived in 2-channel format was also split into two separate streams. This was divided into a 50 minute long development and a 2 hour long evaluation set.

Effect of a VAD in 2-channel recognition

It was found, that in a 2-channel recognition task the long non-speech periods could challenge the speech recognizer. During these periods, various noises or – in most cases – crosstalk was present in the background, which was recognized with high error rate due to the low signal-to-noise ratio. During the first phase of our work, roughly 50% of the training data was transcribed. Though we already added the Supplementary database to the training set, the silence model (trained according to the Baseline approach) was unable to handle these long silences correctly. Therefore, we applied a 2-class VAD, which basically determined the silence boundaries, and recognition was performed on the segments classified as speech. After the transcription work was finished, we repeated this test using the final models – trained on the full Insur and Supplementary dataset with Baseline approach – to determine whether the VAD front-end is still a necessity. The test results obtained on the development set are displayed on Table 6.

Results suggest, that the application of VAD can be helpful in an under-resourced situation. During the early phase of development, it gave more than 30% relative WER reduction on the agent side. However, if we used the final acoustic models, it seemed to have negligible effects. It barely affected the agent side, and caused only 4.5% relative WER reduction on the client side.
Table 6: Tests on the *Insur* development set with First and Final models, without and with VAD (agent/client were left in separate streams)

<table>
<thead>
<tr>
<th>Model</th>
<th>VAD</th>
<th>WER</th>
<th>LER</th>
</tr>
</thead>
<tbody>
<tr>
<td>First, Agent</td>
<td>No</td>
<td>36.7%</td>
<td>24.7%</td>
</tr>
<tr>
<td></td>
<td>Yes</td>
<td>25.6%</td>
<td>11.0%</td>
</tr>
<tr>
<td>First, Client</td>
<td>No</td>
<td>48.8%</td>
<td>30.9%</td>
</tr>
<tr>
<td></td>
<td>Yes</td>
<td>48.4%</td>
<td>30.3%</td>
</tr>
<tr>
<td>Final, Agent</td>
<td>No</td>
<td>15.5%</td>
<td>6.5%</td>
</tr>
<tr>
<td></td>
<td>Yes</td>
<td>15.5%</td>
<td>6.6%</td>
</tr>
<tr>
<td>Final, Client</td>
<td>No</td>
<td>46.3%</td>
<td>28.9%</td>
</tr>
<tr>
<td></td>
<td>Yes</td>
<td>44.2%</td>
<td>28.2%</td>
</tr>
</tbody>
</table>

Comparison of 2-channel and 1-channel models

The main focus of these tests were to determine whether it is necessary to model and recognize the agent and client channel independently, or a single model is sufficient for both channels. Table 7 shows the corresponding results on the evaluation test set. The upper two rows show that case, when models were trained specifically for each channel (2-channel model), and the lower two rows show that case when we only trained a single model (1-channel model). Results show minor differences in the error rates, but these are not significant, thus 1-channel models are well fit for recognizing both audio channels.

To the best of our knowledge, these error rates obtained on the agent channel are the lowest ones achieved up to now on a Hungarian telephone database. Explicit non-verbal event modeling and recognition has not been performed on this database yet, because the majority of the audio data is agent speech, therefore it contains much less non-word expressions.

Recognition of 2-channel and 1-channel audio data

Additionally, it is worth investigating how much can we gain in speech recognition performance by processing the two audio channels separately, compared to the case when we mix them together into a single channel file. Table 8 shows results on the evaluation set for both cases using an out-of-domain model – also trained on insurance data, but different topics, and also using Baseline approach –, and the in-domain 1-channel *Insur* model, which is fit for recognizing the content of both channels. The recognition of 2-channel audio data requires double processing (one for each channel), however results suggest, that high relative improvement (18.1%) can be achieved in the word error rate.

7.2 Personalization and explicit non-verbal modeling tests

In this section the bank call center transcription system is evaluated in various configurations. Due to the smaller size of transcribed database, only a single test set was selected with 2 hours length, instead of separating it to a development and evaluation set. This test data contained both inbound and outbound calls, 75% was originated from the company and 25% was incoming from customers. Our aim was to find out how task specific and additional knowledge sources can help to improve recognition performance.

Effectiveness of model personalization

In the first phase the benefit of utilizing in-domain training data was investigated (see Table 9). The initial recognition results were measured with *Insur* system. First the *Insur* acoustic model was replaced with the task-specific *Bank* model, then only the language model and finally both. As you can see the language model can benefit more of in-domain training data, since WER reduction is almost twice as large if we replace the language model as with the task-specific acoustic model. A probable reason could be that *Insur* and *Bank* tasks have more in common in acoustics than in lexical knowledge.

Table 7: Tests on the 2-channel *Insur* evaluation set using the 2-channel and 1-channel *Insur* models

<table>
<thead>
<tr>
<th><em>Insur</em> Model</th>
<th>Test channel</th>
<th>WER</th>
<th>LER</th>
</tr>
</thead>
<tbody>
<tr>
<td>2-channel, Agent</td>
<td>Agent side</td>
<td>14.7%</td>
<td>7.6%</td>
</tr>
<tr>
<td>2-channel, Client</td>
<td>Client side</td>
<td>44.2%</td>
<td>28.2%</td>
</tr>
<tr>
<td>1-channel</td>
<td>Agent side</td>
<td>14.6%</td>
<td>7.7%</td>
</tr>
<tr>
<td></td>
<td>Client side</td>
<td>45.3%</td>
<td>28.8%</td>
</tr>
</tbody>
</table>

Table 8: Tests on the *Insur* evaluation set with audio channels mixed into a single one, or left separately

<table>
<thead>
<tr>
<th>Model</th>
<th>Audio format</th>
<th>WER</th>
<th>LER</th>
</tr>
</thead>
<tbody>
<tr>
<td>Out-of-domain</td>
<td>Mixed into 1-ch.</td>
<td>59.3%</td>
<td>37.6%</td>
</tr>
<tr>
<td>2 channels</td>
<td></td>
<td>48.8%</td>
<td>25.5%</td>
</tr>
<tr>
<td><em>Insur</em> (1 channel)</td>
<td>Mixed into 1-ch.</td>
<td>33.1%</td>
<td>20.5%</td>
</tr>
<tr>
<td>2 channels</td>
<td></td>
<td>27.1%</td>
<td>13.8%</td>
</tr>
</tbody>
</table>

Table 9: Improvements on *Bank* recognition task by utilizing task specific models

<table>
<thead>
<tr>
<th>Acoustic model</th>
<th>Language model</th>
<th>WER</th>
<th>LER</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Insur</em></td>
<td><em>Insur</em></td>
<td>54.3%</td>
<td>25.7%</td>
</tr>
<tr>
<td><em>Bank</em></td>
<td><em>Insur</em></td>
<td>49.5%</td>
<td>22.1%</td>
</tr>
<tr>
<td><em>Insur</em></td>
<td><em>Bank</em></td>
<td>44.8%</td>
<td>21.2%</td>
</tr>
<tr>
<td><em>Bank</em></td>
<td><em>Bank</em></td>
<td>41.8%</td>
<td>18.9%</td>
</tr>
</tbody>
</table>
and both acoustic models were partially trained on the same audio data, the Supplementary corpus. If manual transcriptions of Bank task are used for training both the acoustic and language model 23% relative WER reduction can be achieved over the system trained for Insur call center task.

### Improvements by knowledge source enhancement

In the next phase of our experiments we focused on the integration of additional textual training data to the baseline bank system. Transcriptions of the Supplementary set, the Insur task, the Bank e-mail corpus and the agent manual are adapted one-by-one to the baseline system by using language model interpolation. As results show in Table 10 additional knowledge source significantly reduces the error rate and altogether resulting in 17% relative improvement in terms of WER and 14% in terms of LER. Models supporting non-verbal event recognition further reduces WER with around 1% due to their ability to detect those acoustical events that would otherwise cause misrecognized words.

All in all, if both language model adaptation and event recognition techniques are considered, 19% relative WER improvement was achieved over the baseline Bank transcription system. Compared to the Insur model, this improvement goes up to 38%. We also estimated the accuracy of the non-verbal event recognition by removing all words from the output, leaving only the series of event labels. This way, the evaluation gave an event recognition accuracy fluctuating around 50% with low variance across the different language models.

Lastly, we expanded the best Bank models by adding the explicit non-speech modeling, while also keeping all the added knowledge sources and non-verbal event models. We have achieved a 35.8% WER and 18.1% LER for speech. Although it gave an almost 6% relative increase in WER compared to the best result, we should not forget, that the output became rich in labels indicating non-verbal and non-speech acoustic content.

### Table 10: Tests on Bank data with using additional knowledge sources, without and with explicit non-verbal event modeling

<table>
<thead>
<tr>
<th></th>
<th>WER Standard</th>
<th>WER Non-v. events</th>
<th>LER Standard</th>
<th>LER Non-v. events</th>
</tr>
</thead>
<tbody>
<tr>
<td>Transcriptions</td>
<td>41.8%</td>
<td>40.8%</td>
<td>18.9%</td>
<td>18.6%</td>
</tr>
<tr>
<td>+Insur and Supplementary</td>
<td>38.1%</td>
<td>37.2%</td>
<td>17.3%</td>
<td>17.0%</td>
</tr>
<tr>
<td>+Email corpus</td>
<td>36.6%</td>
<td>35.6%</td>
<td>16.7%</td>
<td>16.3%</td>
</tr>
<tr>
<td>+Agent manual</td>
<td>34.7%</td>
<td>33.8%</td>
<td>16.2%</td>
<td>15.7%</td>
</tr>
</tbody>
</table>

8 Conclusions

The large scale processing of real-life telephone audio data is a challenging task. As an integral part of a CogInfoCom application, not only hundreds of hours of audio data need to be processed day by day, but also such outputs must be generated, which can support content analysis, data mining, or other kinds of analytic post-processing.

Acoustic and language model training techniques and large vocabulary continuous speech recognition results were presented on two telephone database varying in audio quality, channel number, topic and origin of call. Various approaches were applied to handle the occurring non-speech and speech-like non-verbal acoustic events to increase the application’s reliability and to produce richer recognition outputs.

We primarily used a baseline training approach, where the speaker related and other noises were trained and added to the silence model, thus the system became noise robust. Explicit models were also trained for non-verbal acoustic expressions as well as non-speech events, and their impact on recognition accuracy was investigated. It was found, that the addition of these events comes with speech recognition accuracy improvement in most cases, but most importantly the generated output becomes richer with meta-data regarding the speakers and the recording environment.

Using the two channel Insur database, we found that it has a significant advantage if there is access to audio data, where the speakers are on different channels. 2-channel recognition gives 18% relative improvement over the single channel decoding. It was also found, that it is not necessary to train dedicated models for the separate channels. By training a single acoustical and language model for both channels we achieved nearly the same error rates. Moreover, the 14.6% WER measured on the agent channel is – to the best of our knowledge – the lowest one achieved up to now on a Hungarian telephone database.

The first experiments on Bank database were dedicated to the determine the benefit of using task-specific training data on a call center recognition task. Based on performance tests with various combinations of recognition models, it was found that 23% relative WER improvement can be achieved over the Insur model by applying in-domain training data. The small size of training database left space for further improvement, thus the potentials in integration of additional textual training data were also investigated. Transcriptions of the Insur task and two Bank task specific training corpora were adapted one-by-one to the language model resulting in further 17% relative WER reduction. Detection of non-verbal events further reduced WER with around 1% absolute.

In the future we would like to increase portability of transcriptions across different call center tasks by using...
class-based n-gram language modeling technique for expressions like names, products, etc. Application of special feature sets (e.g. NMF) or neural network-based classifiers could further improve detection rate of non-verbal expressions.

Acknowledgement

Our research was partially funded by the KMOP 1.1.1-07/1-2008-0034 (Knowledge Center), GOP-2009-1.1.1-09/1-20090068 (MonSpeech), KMOP-1.1.3-08/A-2009-0006 (Mindroom), TÁMOP-4.2.2-08/1/KMR-2008-0007(ETOCOM), TÁMOP-4.2.2.C-11/1/KONV-2012-0013 (FuturICT.hu), KMR 12-1-2012-0207 (DIANA) projects and by the Aegon Group.

References


