A CONTEXT-AWARE SPEECH RECOGNITION AND UNDERSTANDING SYSTEM FOR AIR TRAFFIC CONTROL DOMAIN

Youssef Oualil[1]*, Dietrich Klakow*[1]*, György Szaszák*[1]*,*
Ajay Srinivasamurthy[3]*, Hartmut Helmke*[2]*, Petr Motlicek*[3]

[1]Spoken Language Systems Group (LSV), Saarland University (UdS), Saarbrücken, Germany
[2]Institute for Flight Guidance, German Aerospace Center (DLR), Braunschweig, Germany
[3]Idiap Research Institute, Martigny, Switzerland

firstname.lastname@{lsv.uni-saarland.de, dlr.de, idiap.ch}

ABSTRACT

Automatic Speech Recognition and Understanding (ASRU) systems can generally use temporal and situational context information to improve their performance for a given task. This is typically done by rescoring the ASR hypotheses or by dynamically adapting the ASR models. For some domains such as Air Traffic Control (ATC), this context information can be however, small in size, partial and available only as abstract concepts (e.g. airline codes), which are difficult to map into full possible spoken sentences to perform rescoring or adaptation. This paper presents a multi-modal ASRU system, which dynamically integrates partial temporal and situational ATC context information to improve its performance. This is done either by 1) extracting word sequences which carry relevant ATC information from ASR N-best lists and then perform a context-based rescoring on the extracted ATC segments or 2) by a partial adaptation of the language model. Experiments conducted on 4 hours of test data from Prague and Vienna approach showed a relative reduction of the ATC command error rate metric by 30% to 50%.

Index Terms— Automatic speech recognition, context-aware systems, air traffic control, spoken language understanding.

1. INTRODUCTION

Automatic Speech Recognition and Understanding (ASRU) applications can generally benefit from the presence of task-related situational and temporal context (prior) information to improve their performance [1]. This can be done either by 1) refining the ASRU models, such as adapting the acoustic model to new acoustic conditions or adapting the Language Model (LM) to a new domain, or 2) by rescoring the ASR hypotheses using a domain-dependent model. Early usage of situational context goes back to Young et al.'s works [2, 3], who used sets of contextual constraints to generate several grammars for different contexts. Fügen et al. [4] used a dialogue-based context to update a Recursive Transition Network (RTN) to improve ASR quality of a dialogue system. Everitt et al. [5] proposed a dialogue system for gyms, which, based on the exercise routine, would switch its ASR component between pre-existing grammars tailored to different sports equipments.

While there is no doubt that context can significantly improve ASRU performance, the information it provides however, can be small in size, time-varying, partial and available only as machine-generated abstract representations (e.g. airline codes on a radar screen), which are difficult to map back into full possible spoken sentences to perform rescoring or adaptation. In particular, in order to manage a given airspace, Air Traffic Controllers (ATCOs) issue verbal commands to the pilots by interpreting and relying on 1) situational context acquired through multiple modalities such as, radar derived aircraft state vectors comprising position, speed, altitude, etc., as well as 2) temporal context given by the sequence of previously issued commands. Furthermore, verbal communication is the primary mode of communication between agents operating in the ATC domain, which inspires many ASRU-based applications to enhance the ATC technologies. The designed ASRU systems can also benefit from the same context information used by ATCOs. Shore et al. [6] investigated this idea using lattice rescoring on a small Context Free Grammar (CFG)-based simulated ATC setup, whereas Schmidt et al. [7] proposed a dynamic finite state transducer adaptation of a CFG-based LM. As an alternative to CFG solutions, we have recently proposed a Levenshtein-based context integration approach combined with a Statistical Language Model (SLM) [8]. More details about ASRU for ATC are presented in Section 2.

This paper extends and generalizes the work presented in [7, 8] in different directions. That is, 1) in addition to situational context, we propose a new model that also integrates temporal context (history of spoken commands) (Section 3). Then, 2) we combine the two types of context in a generalization of [8] using N-best lists (Section 4). Finally, 3) contrary to [7, 8], which evaluated their systems on data collected from a simulator of Düsseldorf airport, this paper evaluates the system on 4 hours of data collected from ATCOs performing their daily tasks in Vienna and Prague airports (Section 6). The obtained results show that the proposed context-aware ASRU system reduces the ATC Command Error Rate (CmdER) metric by 30% to 50% compared to a standard ASRU system.

2. ASRU SYSTEMS FOR ATC DOMAIN

2.1. Air Traffic Control Assistance Systems

The task of air traffic control aims at maintaining a safe, orderly and expeditious flow of air traffic. ATCOs apply strict separation rules to direct aircraft safely and efficiently, both in their respective airspace sector and on the ground. Since controllers have a significant responsibility and can face high workloads in busy sectors, different planning systems have been proposed to assist them in managing the airspace such as, the Arrival Manager (AMAN). These systems mainly suggest an optimal sequence of commands (command advisories), which are then issued in verbal radio communication from the controller to the aircraft pilots.

This work was supported by the MALORCA project (Grant Agreement No. 698824), funded by SESAR Joint Undertaking, under EU Horizon 2020.

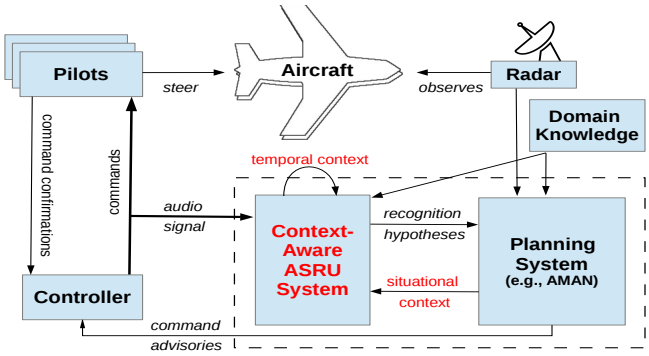

Fig. 1. Schematic view of an ASRU-based ATC system.

2.2. AcListant®: Active Listening Assistance System

For different reasons such as, emergency or weather conditions, the controller may deviate from the advisory commands proposed by the assistance system. The latter reacts slowly to such deviations and may require the controller to enter the issued commands via mouse/keyboard. Thus, indirectly increasing the workload that they were mainly designed to reduce. As a solution to this problem, we have recently proposed the AcListant®[1] [9] system, which extends the planner to include a background ASRU system, ideally replacing the mouse/keyboard feedback. Conversely, ASRU can also benefit from the context information used by the assistant system [8, 10] to improve its performance. We will refer to it as Assistance-based ASRU (ABSRU) system in the rest of this paper. Fig. 1 shows the information flow in an ASRU-based assistance system.

2.3. From AcListant® To MALORCA

Although the AcListant® system achieved a good performance in a simulator of Düsseldorf airport, the cost of transferring such system from the laboratory to real ops-rooms is very significant. Each model in the ABSRU system must by manually adapted to the linguistic and acoustic features of the new environment, which are due to new local conditions such as, noise conditions, different accents, speaking styles, deviations from standard phraseology [11], etc. Therefore, the MALORCA[2] project is proposed as a generalization of AcListant® that aims at developing a general, cheap and effective solution to automate the re-learning, adaptation and customization process to new environments. This will be done by taking advantage of the large amount of un-transcribed speech data available on a daily basis in the new ATC environment, which can be used in un/semi-supervised learning approaches to automatically adapt the ABSRU models to the respective environment. The work presented in this paper describes the basic and general ABSRU systems, which will be used as initial points in the bootstrap automatic adaptation cycle for Vienna and Prague airports, respectively.

3. ATC CONTEXT-BASED RESCORING

This section introduces the different types of context we consider and the mathematical models we designed to integrate them into an ASRU system. Then, we show how these different models can be combined in a unifying framework.

[1]AcListant®: http://www.AcListant.de

[2]MALORCA: MAchine Learning Of speech Recognition models for Controller Assistance: http://www.malorca-project.de

3.1. Situational Context Information

An ATC assistance system bases its proposed command sequence on the state of a given airspace sector. This state is primarily derived from radar information about the current situation of the airspace and aviation domain knowledge. This is done by forming a search space of all physically possible commands in the current airspace situation in a first step, and then extracting the advisory sequence of commands, shown to ATCOs, by optimizing a set of ATC criteria. The formed search space summarizes the current situation in the airspace. Thus, we will refer to it as **situational context**. For an ASRU system, this context can be seen as a command-level search space, which is 1) dynamic, i.e. changes every few seconds, 2) small in size, i.e. few hundred/thousand of commands, and 3) available only as partial standardized ICAO phraseology concepts [11] (see example Table 1). In particular, a situational context information contains an aircraft callsign (e.g. AFR2A ≅ *air france two alpha*) followed by a command type to execute and a command value to achieve (e.g. REDUCE 220 ≅ *reduce speed two two zero knots*).

Callsign	Command Type	Value
AFR2A	REDUCE	220
DLH9000	DESCEND	120
BER256	RATE_OF_DESCENT	3000
KLM23RV	TURN_LEFT_HEADING	80

Table 1. Excerpt from situational context information generated by a planning system. It shows an ICAO abstraction of four different actions that can be issued by the controller to an aircraft.

Given the spoken language variability, it is very difficult to build the word-level context space, which maps each command in the context into the set of all possible spoken realizations of that command, which can be issued by an ATCO to an aircraft pilot. Furthermore, such process should be very fast given that the situational context changes every few seconds. As a result, performing the standard lattice rescoring or LM adaptation is not feasible in this case. The next section introduces a partial rescoring approach, which considers only the ATC segments in the recognized hypotheses.

3.2. Situational Context-based Rescoring (SCR)

The situational context model considers the context information as an ASRU search space for ATC concepts. That is, it only targets sequence of words that carry some ATC information in the recognized hypotheses. This partial rescoring approach follows these steps:

Step 1) Sequence Labeling: This step takes the raw ASR hypothesis as input and automatically **detects and extracts** the ATC concepts that it carries. For instance, the hypothesis "*air france two alpha hello reduce speed two three zero knots*" is mapped to "`<callsign>` *air france two alpha* `</callsign>` *hello* `<command=reduce>` *reduce speed* `<speed>` *two three zero* `</speed>` *knots* `</command>`". This step directly puts the focus on the ATC information carried by the ASR hypotheses, which is our primary target, and ignores the rest. Our experiments use a CFG-based token tagger similar to the one used in [7, 8].

Step 2) Context-to-Word Mapping: The partial rescoring approach turns the problem of generating full spoken sentences (realizations) of the context into generating realization of short segments, which can be extracted by the sequence labeler in the previous step. For instance, instead of generating the full realization of the command "AFR2A REDUCE 250", we only need to generate context-to-word mapping for the callsign "AFR2A" and the speed value "250".

Step 3) Situational Context-based Rescoring: We use here a Weighted Levenshtein Distance (WLD) to rescore the ATC segments extracted from the ASR hypotheses in Step 1, in the search space formed by all verbalized context segments from Step 2. More details about the WLD can be found in [8].

Formally, let $A = \{A_{cs}, \{A^{cs}_{com}\}_{com}\}$ be the ATC segments extracted from the ASR hypothesis using sequence labeling as described in Step 1. We assume that each hypothesis contains (at most) a single callsign A_{cs} in addition to one or multiple issued commands $\{A^{cs}_{com}\}_{com}$. Similarly, let $\mathcal{C} = \cup_{cs}\{(C_{cs}, \{C^{cs}_{com}\}_{com})\}$ be the set of all possible context-based ground truths resulting from the context-to-word mapping described in Step 2. This set consists of all callsigns in the context and the ATC commands applicable to them. The situational context-based rescoring extracts the "corrected" ATC segments $H = \{H_{cs}, \{H^{cs}_{com}\}_{com}\}$ according to

$$\begin{aligned} H &= \underset{C\in\mathcal{C}}{\operatorname{argmin}}\{WLD(A, C)\} \qquad (1)\\ &= \underset{C\in\mathcal{C}}{\operatorname{argmin}}\{WLD(A_{cs}, C_{cs}) + \sum_{A_k\in\{A^{cs}_{com}\}} \underset{C_j\in\{C^{cs}_{com}\}}{\operatorname{argmin}} WLD(A_k, C_j)\} \end{aligned}$$

More details about the WLD and the situational context-based rescoring can be found in [8].

3.3. Temporal Context-based Rescoring (TCR)

Air traffic control assistance systems typically use the radar information to generate the situational context. The resulting command advisories are generated through a deterministic optimization process, which takes into account a number of physical and local constraints about the operating airport. These constraints include waypoints, which play the role of "markers" in the airspace, location of the runways for landing and departure, the landscape surrounding the airport (see, mountains, etc), to name a few. Due to these constraints, a number of pre-defined trajectories and landing patterns are frequently generated to guide aircraft from their current location to the runways. For instance, most landing aircrafts will receive a confirmation of identification as first command, and a handover as last command. In particular, once an aircraft enters the controlled airspace, the generated landing sequence for this aircraft is expected to be closely similar to the ones generated for previous aircraft that entered that airspace at close locations. Fig. 2 shows an example of landing sequence and trajectory patterns that are expected to be followed by different aircraft depending on their location.

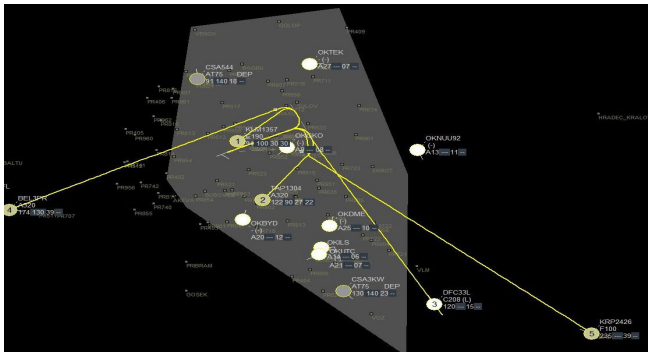

Fig. 2. Expected landing sequences and trajectories for different aircraft approaching Prague airport.

Based on these pre-defined patterns, we designed an "Airport Flight Model", which can predict the future commands to be spoken to a given aircraft based on the **history** (temporal context) of the previously issued commands to that aircraft.

In practice, this model is a Long-Short Term Memory (LSTM) neural network [12, 13] trained on landing sequences of commands, which are reconstructed from data collected in Prague or Vienna airports. That is, we define the input to this model as the timely-ordered sequence of commands, which were issued to a given aircraft since it entered the controlled airspace and until it landed on the runway. The next section shows how this temporal model can be combined with the situational model to generalize the approach proposed in [13].

4. A GENERALIZED CONTEXT-AWARE ASRU SYSTEM

Although the SCR approach (Section 3.2) can significantly improve the performance, it only operates on command values and callsigns. More precisely, if the ASRU hypothesis confuses two commands which take the same attribute but are of different types, the SCR will not be able to correct this misrecognition. e.g. the sequence labeler extracts a "SPEED 220" command instead of a "REDUCE 230", which both take a speed value as attribute. In this case, SCR would be able to correct the command value "220" to "230" but cannot correct the command type "SPEED" to "REDUCE".

This problem can be solved using the TCR approach (Section 3.3). In order to do so, we train this model only on command types without command values, i.e. we only predict the probability of a "REDUCE" command in a given context regardless of the speed value that can be assigned to it. This in fact is a marginalization of the full model (command type+value) on the complete range of values that this command can take. Furthermore, this decision is also justified by the small amount of data available to train the full model, which would result in a vocabulary size of few hundred/thousand, resulting from the rich range of values that each command can take. Building a model only for command types reduces drastically the vocabulary size (40 to 60 different command types).

In order to combine the SCR and TCR models, we consider N-best lists instead of 1-best hypothesis which was used in [8]. Formally, assuming the ASR system produces a list of N hypotheses, let $A = \{A^n\}_{n=1}^N = \{\{A^n_{cs}, \{A^{cs,n}_{com}\}_{com}\}\}_{n=1}^N$ be the set of ATC segments extracted from these hypotheses using sequence labeling (Section 3.2). The combination of SCR and TCR models is done according to

$$\begin{aligned} H &= \underset{n=1,\ldots,N}{\operatorname{argmin}}\{\underset{C\in\mathcal{C}}{\operatorname{argmin}}\{p(A^n, C)\}\} \qquad (2)\\ &= \underset{n=1,\ldots,N}{\operatorname{argmin}}\Big\{\underset{C\in\mathcal{C}}{\operatorname{argmin}}\{p^s(A^n_{cs}, C_{cs})\}\\ &+ \sum_{A^n_k\in\{A^{cs,n}_{com}\}} \underset{C_j\in\{C^{cs}_{com}\}}{\operatorname{argmin}} \{p^t_{cs}(A^n_k)\cdot p^s(A^n_k, C_j)\}\Big\} \end{aligned}$$

The probability $p(A^n, C)$ combines 1) a situational context based-rescoring probability $p^s(.,.)$, directly derived from the WLD scores used in Section 3.2. This distribution operates on callsigns and command values as explained above, and 2) a temporal context based score $p^t_{cs}(.)$, which estimates the probability distribution over the command type space given the history of issued commands for a callsign cs. In doing so, the situational and temporal context models complement each other, which leads to a generalized model that can successfully rescore callsigns, command types and command values.

5. PRAGUE AND VIENNA DATASETS

The proposed context-based rescoring system is evaluated using recordings of actual ATCOs performing their daily tasks in Prague and Vienna airports. This data was collected as part of the MALORCA[2] project. It consists of 8kHz ATC speech recordings of different noise levels and different radio transmission qualities. In par-

ASRU Systems	Prague Results (Error Rates are in %)					Vienna Results (Error Rates are in %)				
	WER	ConER	CmdER	$\overline{\text{CmdER}}$	R_t(s)	WER	ConER	CmdER	$\overline{\text{CmdER}}$	R_t(s)
SLM (no context)	10.9	17.5	30.9	21.9	1.25	**13.2**	22.3	41.4	30.4	0.90
SLM+Rescoring (N-best=1)	11.2	13.8	19.1	12.8	3.40	17.5	16.4	27.7	20.6	3.22
SLM+Rescoring (N-best=5)	**8.9**	**11.6**	**16.5**	**12.7**	4.65	15.5	**15.5**	**26.3**	**19.8**	3.63
CFG (no context)	18.0	33.1	50.5	37.5	**1.02**	22.1	38.5	58.9	43.1	**0.77**
CFG+Adaptation	17.8	21.9	30.9	23.4	3.57	26.7	29.7	44.1	30.4	1.43
CFG+Rescoring (N-best=1)	19.7	25.3	33.1	25.4	1.87	25.6	27.9	40.1	33.0	1.71
CFG+Rescoring (N-best=10)	19.1	24.4	31.8	24.2	4.57	25.1	26.5	38.5	32.0	2.97

Table 2. ASRU results on 4h of test data from Prague and Vienna airports using different ASRU systems with and without context information.

ticular, the Vienna dataset is very noisy and it can be difficult to understand for humans with no ATC expertise. All commands were issued in English with a mild usage of Czech or Austrian German languages, respectively. In particular for words which do not contain any ATC information such as greetings. Different ATC sessions were recorded over multiple days from each controller. Table 3 presents recording statistics for these two datasets.

The situational context is updated every 5 seconds by the assistant system [10]. Table 3 also reports the context accuracy, i.e. context contains the actual spoken command, and the average context size i.e. number of ATC commands per context file, which can be compared to 239 and 359 used in [7] and [8], respectively.

	Duration (h)		# of Speakers		Context	
	Train	Test	Train	Test	Size	Acc.
Prague	2.1h	1.9h	6	5	650	99.0%
Vienna	5.0h	1.9h	13	6	1600	96.0%

Table 3. Recording statistics for Vienna and Prague datasets including the context accuracy (i.e. contains the actual spoken commands).

6. EXPERIMENTAL SETUP AND ANALYSIS

ASR was performed using the KALDI software [14] and the ASR confidence scores for WLD were generated based on the Minimum Bayesian Risk (MBR) decoding approach [15]. The acoustic model is a DNN/HMM (Deep Neural Network Hidden Markov Model), trained on 150 hours of speech data from the publicly available LIBRISPEECH [16], ICSI [17], AMI [18] and TED-LIUM [19] datasets, which have been extensively used in ASR of conversational speech, and then adapted on Vienna or Prague training data in Table 3. More details about this system can be found in [20]. The SLM is a trigram model trained on a combination of the training data and synthetic data generated from the CFG. The latter defines its rules based on the standard ATC phraseology [11], in addition to most common deviations observed in the training data. The CFG design was guided by the work done in [7, 8].

For evaluation, in addition to conventional WER and Recognition time (R_t), the ATC-specific evaluation metrics *Concept Error Rate* (ConER) and CmdER are used. ConER is restricted to the ATC-relevant semantic concepts of a given utterance, which are extracted using the sequence labeling approach (Section 3.2). A concept can be either a callsign or a command, e.g. AFR2A or REDUCE_250. The CmdER metric requires the entire sequence of concepts to be correct. In the case where the sequence labeling system fails in extracting ATC segments, it returns NO_CALLSIGN or NO_COMMAND, which are counted as misrecognition, even though they have no impact on the planning system (no information). Therefore, we also report the CmdER after excluding these utterances (noted $\overline{\text{CmdER}}$) to estimate the misrecognition rate which negatively affects the planning system.

Table 2 reports the ASRU results for Vienna and Prague test data with and without context information. The approach "CFG+Adaptation" is the one proposed in [7]. Furthermore, using an N-best=1 is equivalent to the system proposed in [8], which does not use temporal context. In this case, the recognized ATC segment contains (at most) one command type. Thus, the TCR is not used.

The results clearly confirm the conclusions reported in [8]. That is, SLM clearly outperforms the CFG-based system with and without context information. This observation highlights the importance of the probability distribution learned by SLM but ignored by CFG, which uses a uniform distribution over words and commands. Moreover, SLM automatically captures deviations from standard phraseology present in the data, whereas CFG requires a manual addition.

We can also conclude from these results that context information strongly improves the ATC-related metrics (ConER, CmdER and $\overline{\text{CmdER}}$), whereas it slightly improves or worsens the WER of either system. This is an expected outcome given that the proposed approach is mainly designed to improve the ConER (and therefore also the CmdER), by directly extracting and correcting ATC segments from the recognized hypotheses. Correcting such segments, however, does not necessarily mean improving the word-level recognition. This is particularly true in cases where the controller deviates from standard phraseology [11], which was used to build the context-to-word mapping (Section 3.2), e.g. dropping the word "decimal" while issuing the frequency 133.2 ="one three three decimal two". These cases were very common, particularly in Vienna data. Furthermore, increasing the N-best list size leads to further improvements for all systems. This observation highlights the advantages of the proposed generalized system compared to the one proposed in [8] (N-best=1). In fact, testing the TCR component alone leads to an accuracy (prediction of the command type) of 59% for Prague and 55% for Vienna, with a mean rank of 2.4 and 2.7, respectively.

These experiments also show that data and context quality are very crucial. More particularly, the Prague speech data is less noisy compared to Vienna data and largely benefits from the smaller and more accurate situational context (Table 3). Moreover, comparing CmdER and $\overline{\text{CmdER}}$ shows an average degradation of $\approx 10\%$. This reflects the need for a better sequence labeler to extract the ATC segments. The recognition time R_t, however, is within a real-time range given that ATC utterances are $\approx 3.7s$ long on average.

7. CONCLUSIONS AND FUTURE WORK

We proposed a context-aware ASRU system for ATC domain, which combines situational context acquired through an ATC assistance system, and temporal context given by the history of issued commands. Experiments conducted on real data from Prague and Vienna airports showed a significant reduction of the command error rate. Our future work will focus on investigating different sequence labeling approaches, which seem to be a cornerstone for improving the performance of the overall system.

8. REFERENCES

[1] Geert-Jan M. Kruijff, Pierre Lison, Trevor Benjamin, Henrik Jacobsson, Hendrik Zender, and Ivana Kruijff-Korbayová, "Situated dialogue processing for human-robot interaction," in *Cognitive Systems*, vol. 8 of *Cognitive Systems Monographs*, chapter 8, pp. 311–364. Springer Verlag, Berlin/Heidelberg, Germany, 2010.

[2] Sheryl R. Young, Wayne H. Ward, and Alexander G. Hauptmann, "Layering predictions: Flexible use of dialog expectation in speech recognition," in *Proceedings of the 11th International Joint Conference on Artificial Intelligence. Detroit, MI, USA, August 1989*, 1989, pp. 1543–1549.

[3] S. R. Young, A. G. Hauptmann, W. H. Ward, E. T. Smith, and P. Werner, "High level knowledge sources in usable speech recognition systems," *Commun. ACM*, vol. 32, no. 2, pp. 183–194, Feb. 1989.

[4] Christian Fügen, Hartwig Holzapfel, and Alex Waibel, "Tight coupling of speech recognition and dialog management - dialog-context dependent grammar weighting for speech recognition," in *INTERSPEECH 2004 - ICSLP, 8th International Conference on Spoken Language Processing, Jeju Island, Korea, October 4-8, 2004*.

[5] Katherine Everitt, Susumu Harada, Jeff A. Bilmes, and James A. Landay, "Disambiguating speech commands using physical context," in *Proceedings of the 9th International Conference on Multimodal Interfaces, ICMI 2007, Nagoya, Aichi, Japan, November 12-15, 2007*, 2007, pp. 247–254.

[6] Todd Shore, Friedrich Faubel, Hartmut Helmke, and Dietrich Klakow, "Knowledge-based word lattice rescoring in a dynamic context," in *INTERSPEECH 2012, 13th Annual Conference of the International Speech Communication Association, Portland, Oregon, USA, September 9-13, 2012*, 2012, pp. 1083–1086.

[7] Anna Schmidt, Youssef Oualil, Oliver Ohneiser, Matthias Kleinert, Marc Schulder, Arif Khan, and Hartmut Helmke, "Context-based recognition network adaptation for improving on-line asr in air traffic control," in *2014 IEEE Spoken Language Technology Workshop (SLT 2014)*, 2014, pp. 2–6.

[8] Youssef Oualil, Marc Schulder, Hartmut Helmke, Anna Schmidt, and Dietrich Klakow, "Real-time integration of dynamic context information for improving automatic speech recognition," in *INTERSPEECH 2015, 16th Annual Conference of the International Speech Communication Association, Dresden, Germany, September 6-10, 2015*, 2015, pp. 2107–2111.

[9] Hartmut Helmke, Youssef Oualil, Jürgen Rataj, Thorsten Mühlhausen, Oliver Ohneiser, Heiko Ehr, Matthias Kleinert, and Marc Schulder, "Assistant-based speech recognition for ATM applications," in *Proceedings of* 11^{th} *USA/Europe ATM R&D Seminar (ATM2015)*, Lisbon, Portugal, June 2015.

[10] Hartmut Helmke, Ronny Hann, Maria Uebbing-Rumke, Dennis Müller, and Dennis Wittkowski, "Time-based arrival management for dual threshold operation and continuous descent approaches," in *Proceedings of* 8^{th} *USA/Europe ATM R&D Seminar (ATM2009)*, Napa, California, USA, June - July 2009.

[11] "All clear phraseology manual," in *Eurocontrol, Brussels, Belgium*, April 2011.

[12] Martin Sundermeyer, Ralf Schlüter, and Hermann Ney, "LSTM neural networks for language modeling," in *13th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, Portland, OR, USA, Sep. 2012, pp. 194–197.

[13] Y. Oualil and D. Klakow, "A neural network approach for mixing language models," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 2017, pp. 5710–5714.

[14] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, Jan Silovsky, Georg Stemmer, and Karel Vesely, "The Kaldi speech recognition toolkit," in *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. Dec. 2011, IEEE Signal Processing Society.

[15] Vaibhava Goel and William J. Byrne, "Minimum bayes-risk automatic speech recognition," *Computer Speech & Language*, vol. 14, no. 2, pp. 115–135, 2000.

[16] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur, "Librispeech: an ASR corpus based on public domain audio books," in *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 5206–5210.

[17] Adam Janin, Don Baron, Jane Edwards, Dan Ellis, David Gelbart, Nelson Morgan, Barbara Peskin, Thilo Pfau, Elizabeth Shriberg, Andreas Stolcke, and Chuck Wooters, "The ICSI meeting corpus," in *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. IEEE, 2003.

[18] Jean Carletta, "Announcing the AMI meeting corpus," *The ELRA Newsletter*, vol. 11, no. 1, pp. 3–5, 2006.

[19] Anthony Rousseau, Paul Deléglise, and Yannick Estève, "Enhancing the TED-LIUM Corpus with Selected Data for Language Modeling and More TED Talks," in *Proc. of the 9th International Conference on Language Resources and Evaluation (LREC)*, 2014, pp. 3935–3939.

[20] Ajay Srinivasamurthy, Petr Motlicek, Ivan Himawan, Gyorgy Szaszak, Youssef Oualil, and Hartmut Helmke, "Semi-supervised learning with semantic knowledge extraction for improved speech recognition in air traffic control," in *INTERSPEECH 2017, 18th Annual Conference of the International Speech Communication Association*, Aug. 2017.