DESIGN AND DEPLOYMENT OF MULTILINGUAL INDUSTRIAL VOICE CONTROL APPLICATIONS

I. Kraljevski, M. Pohl, A. Gjoreski, U. Koloska, J. Wöhl, M. Wenzel, D. Hirschfeld

voice INTER connect GmbH, Ammonstraße 35, D-01067 Dresden, Germany
{ivan.kraljevski, matthias.pohl, aleksandar.gjoreski, uwe.koloska, jonas.woehl, martin.wenzel, diane.hirschfeld}@voiceinterconnect.de

Abstract: This paper presents a concept and a demonstrator for rapid design and deployment of multilingual industrial voice control speech applications. The dialogues are modelled in a designer toolkit and the resources are deployed on a target embedded system. The workflow consists of distinctive design and deployment phases: definition of speech interaction by giving representative utterance examples (language modelling); assignment of the control functions (intents or tasks); incorporating domain knowledge (entity definition and interfaces), definition and assignment of the acoustic feedback (text-to-speech); validation, optimization of the speech dialogue; generation of language specific resources; deployment to the target platform; iterative testing and optimizations. The demonstrator presents hands-free and natural language interactions with arbitrary external systems via the standard Internet of Things protocols. It does not require internet access, the speech and audio feedback processing is performed offline, ensuring privacy by design. The presented approach allows rapid prototyping of multilingual dialogues, as a part of an iterative process of assessing and improving the quality of interaction and the user experience.

1 Introduction

Recent advances in speech technology have enabled human-machine interfaces to provide reasonably good spontaneous and unrestricted interactions over a range of devices. Commercial products like digital personal assistants (DPA) and special home devices with a DPA function (smart speakers) are more and more present in the consumers' households and in everyday life [1-2].

In the light of the ongoing revolution of industrial production, worldwide initiatives ("Industry 4.0" in Germany, "Made-in-China 2025", "New Robot Strategy" in Japan, "New Industrial France", "Re-Industrialization" in the USA), took the advantage over continuous innovations where Cyber-Physical Systems (CPS), the Internet of Things (IoT), the Internet of Services (IoS), robotics, big data, cloud and cognitive computing and augmented reality (AR) result in a significant change in production systems [3].

Nevertheless to say, there is a growing trend of deploying industrial voice applications that has been developed to help the operators in a human-like interaction by addressing a variety of tasks: extension of existing graphical user interfaces (GUI) with voice control, control of machines and appliances, operation and monitoring of medical systems, intuitive input in rehabilitation technology, natural collaboration with robots and in augmented reality (AR) applications, self-explanatory user interfaces [4].

There is an increasing importance of global presence which introduces also an increasing number of skilled non-native employees in multinational corporations. Therefore, deploying multilingual industrial voice applications allows those operators actively and equally to contribute to the production process.

Since the voice data is unique in its historical protection, communicative content, and bio-metric features, it raises privacy implication concerns about microphone-enabled devices [5]. In order to achieve the best possible performance of the speech services, sensitive user information (voice, personal preferences) is transmitted to the cloud. However, in many application areas using the cloud for speech services is not possible due to technical limitations or undesirable due to privacy issues. Therefore, implementing high quality human-machine speech interaction on embedded devices is still a challenging task. To provide an acceptable user experience, such multilingual speech interface has to employ robust speech recognition with low computational and memory footprint and it would be used in outdoor and indoor spaces with a significant presence of background noise. The speakers are interacting mostly using spontaneous speech, so the system has to be able to identify complex surface expressions and map them with a particular meaning while handling miscommunication by adaptive dialogue management, using confirmations and recovery strategies.

This paper presents a concept and a demonstrator for rapid design and deployment of multilingual industrial voice control speech applications. In section 2 the challenges of using embedded speech applications are discussed. Section 3 gives an overview of the phases of the speech application design and section 4 presents the use case of an industrial voice control application. The conclusions are given in section 5.

2 Embedded Speech Applications

Implementing Natural Language Understanding (NLU) functionality and quality of interaction on an embedded device is a challenging task due to limited computational resources. To provide an acceptable user experience, the speech interface has to employ robust speech recognition with low computational and memory footprint.

Not long ago, in order to deploy speech recognition on an embedded system the main challenge was not the recognizer itself but the available hardware where it should run. Currently, mobile devices and mini PCs are powerful enough to run automatic speech recognition. But, the variety of such devices prevent to design low cost speech recognition software running on all of them with the same performance level. It is much easier to design and implement simple speech interface which will record and transmit the speech signal to remote speech recognition service for processing and providing back the recognition results.

However, such service is not suitable for many industrial speech applications. Freely available remote speech services are not tailored to a specific application domain. The backend of such remote service is often a general purpose Large Vocabulary Continuous Speech Recognition (LVCSR) system, and because of its size, the recognition accuracy is usually lower than as used with limited domain vocabulary. If used on a limited domain (e.g. 100 words), it will provide worse performance than a system built around limited vocabulary. In this case, it is better to use Context-Free-Grammars (CFG) instead and the accuracy will be higher. Another disadvantage of LVCSR against CFG grammars is that they need a semantic parser and interpreter. Grammars with semantic tags in the same domain will outperform comparable LVCSR systems in terms of accuracy, latency, and required resources.

On the other hand, writing a complex and robust grammar capable to cover spontaneous and naturally formulated speech, even on a limited domain, is very difficult and it restricts the application design to trained and skilled persons. One solution, employed in the presented concept is to model the target domain by word-class Statistical Language Models (SLM). The word classes contain either list of domain specific words or they can be modelled by CFG grammars to handle specific entities.

Latency is another issue, for many systems real-time processing is of the highest importance. The latency depends on the language models size, the used recognition technology, and its configuration. Also transmission between the system and the remote service contributes to the latency, where it could vary between 100 milliseconds and several seconds, and in many cases, the availability of the remote service is not guaranteed.

In different scenarios, the speech applications could be used in outdoor and indoor spaces with a significant presence of background noise, where the users are interacting using spontaneous speech (Lombard effect, hyper-articulation, hesitations, breathing, etc.). Therefore, the system has to be able to handle possible miscommunication by employing adaptive management, using confirmations and recovery strategies. Moreover, the development of multilingual speech applications requires effort, resources, time and expertise [6]. In order to ensure equal usability of the speech interface across languages, the dialogue design process can be performed in parallel for more than one language, considering the basic interaction as language independent.

3 Speech Application Design

As seen in many studies, as well as in our own staged WOz experiments [7], speakers prefer to interact with devices using unrestricted spontaneous speech. To provide a high level of NLU interaction, the embedded speech application has to be able to identify complex surface expressions and map them to a particular meaning. Such functionality is not possible to achieve only by employing handcrafted context-free grammars.

However, a statistical language model of an in-domain text corpus cannot cover all the possible utterance variations (flowery phrases, slot synonyms, etc.) without substantial effort for adaptation. An additional problem is the creation of syntax parser and interpreter to provide a recognition result containing the user intention and its corresponding functional entities (slots). The common approach is to use the aforementioned word-class language modelling where the classes correspond to slots as defined in the dialogue specification. A word class could be described not only by a list of elements but also by a more complex definition, like CFGs or large lists of hierarchically dependent fields [8]. A general overview of the steps to create NLU enabled embedded voice control:

1. Assignment to control functions (semantics);
2. Definition of voice commands (words and phrases);
3. Specification of device information and system feedback;
4. Definition and assignment of the acoustic feedback;
5. Validation, optimization of the speech dialogue;
6. Generation of language resources;
7. Deployment of resources to the target platform;
8. Iterative testing and optimizations;

3.1 Design, deployment, and execution

Figure 1 depicts the complete workflow starting from the initial dialogue design and ending with the deployment of the compiled speech application on the target embedded platform. It consists of a web technology based development environment (frontend), a cloud application (backend) and the embedded system.

The target system architecture is composed of several components. The audio input/output system, the dialogue manager, the NLU module and the control application, all of them interfacing using the MQTT protocol. The component sends data with a predefined topic to the MQTT broker which broadcasts it further to other subscribed components. The dialogue manager is also handling the incoming messages for the subscribed events. In the case of speech interaction, it will start the dialogue to collect information and broadcast back a message with the speaker’s intention and corresponding slot values.

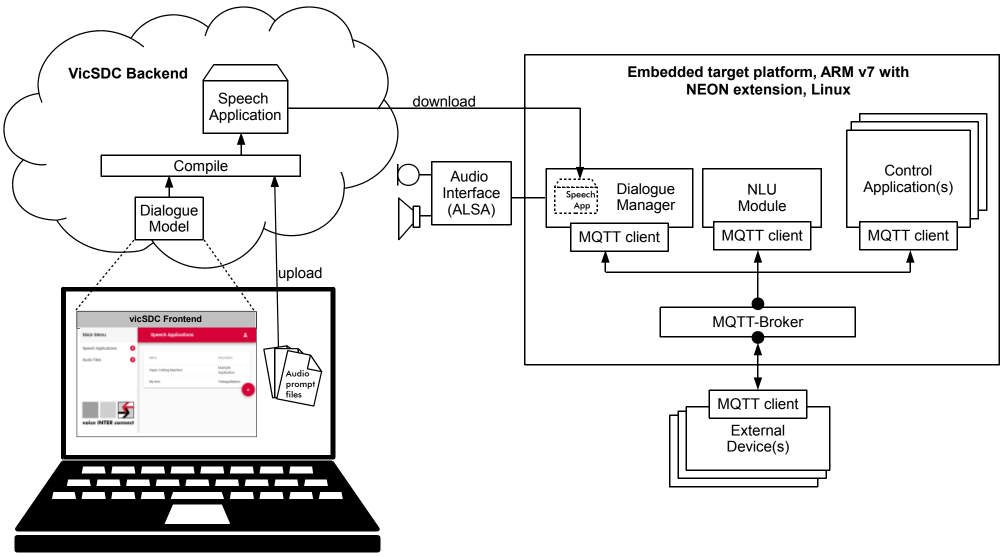
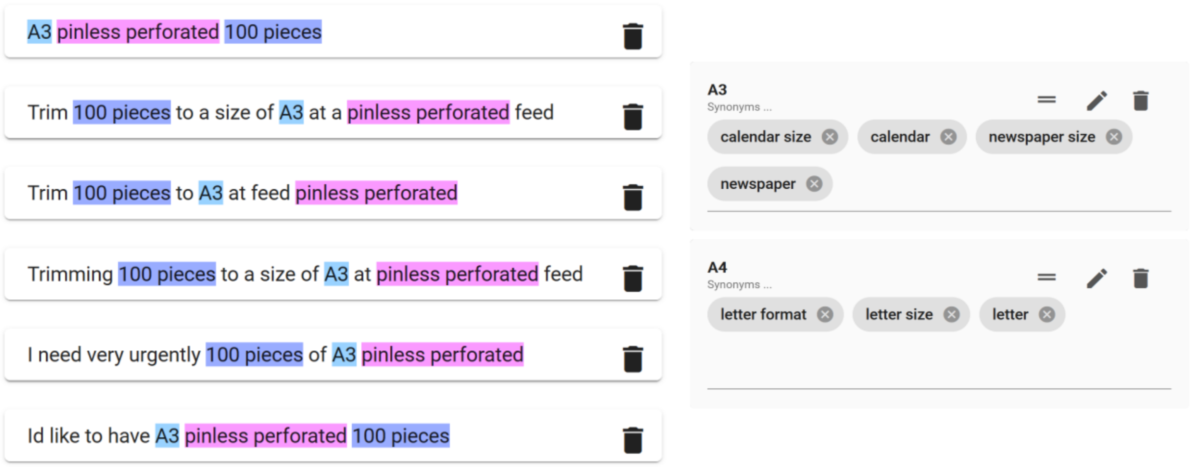


Figure 1 - Design and deployment of industrial voice control applications

3.1.1 Design phase

The dialogue is designed using the web-based tool for speech dialog creation (vicSDC). The designer describes application functionality by defining intents, representing the tasks to be fulfilled, and providing speech interaction examples. Each functionality could have additional arguments that have to be considered in order to complete the task. Therefore, intents could have none or several slots, corresponding to the functional arguments or parameters. The slots are defined as a list of elements where each has a language specific canonical form and optionally a list of synonyms. Alternatively, a slot could be defined by external pre-built grammars with more complex structure, e.g. handling numbers, time and date, etc.

The designer enters representative utterance examples reflecting the way how particular functionality will be used in the speech interaction. The words which are representing the arguments are marked with the corresponding slots (Figure 2). The training sentences along with the slot definitions serve as input for automatic compilation of the language resources: in-domain textual corpus, word class lists, generated and prebuilt external grammars which are the basis for the building of the SLM and the NLU models.



A3 pinless perforated 100 pieces

Trim 100 pieces to a size of A3 at a pinless perforated feed

Trim 100 pieces to A3 at feed pinless perforated

Trimming 100 pieces to a size of A3 at pinless perforated feed

I need very urgently 100 pieces of A3 pinless perforated

Id like to have A3 pinless perforated 100 pieces

A3
Synonyms ...
calendar size ⊗ calendar ⊗ newspaper size ⊗ newspaper ⊗

A4
Synonyms ...
letter format ⊗ letter size ⊗ letter ⊗

Figure 2 – Examples of marked examples sentences (left) and slot values with synonyms (right)

Additionally, specific audio feedback can be assigned to an intent, which would be automatically presented to the speaker as a confirmation after successful recognition and execution. Additional audio samples could be included within the speech application where they can be triggered outside by sending custom formatted MQTT message to the dialogue manager.

3.1.2 Deployment phase

At this stage, the dialogue designer can check the correctness of the model by validating it in the web-tool, in terms of requirements that should be fulfilled as, e.g.: the minimum number of utterances per intent to ensure necessary amount of training data for the machine learning algorithms, missing wake-up-phrase etc.

By starting the compilation phase, the dialogue model specification along with the audio prompts are sent to the backend where all the necessary language and configuration resources are automatically generated and provided for download in a form of a compressed file. The file is transferred to the target platform and placed on the file system where the dialog manager could access it at run-time.

3.1.3 Run-time phase

At run-time, the dialogue manager loads the configuration and language resources of the speech application. This approach has more advantages than compiling monolith applications since it allows shorter interactive design-deploy cycles, faster tests, and optimization. A wake-up-phrase is used to activate the dialogue and start the recognition. After successful recognition, the result is sent over the MQTT interface to the NLU module which classifies the orthography into one of the intents and extracts the slot values if any are present.

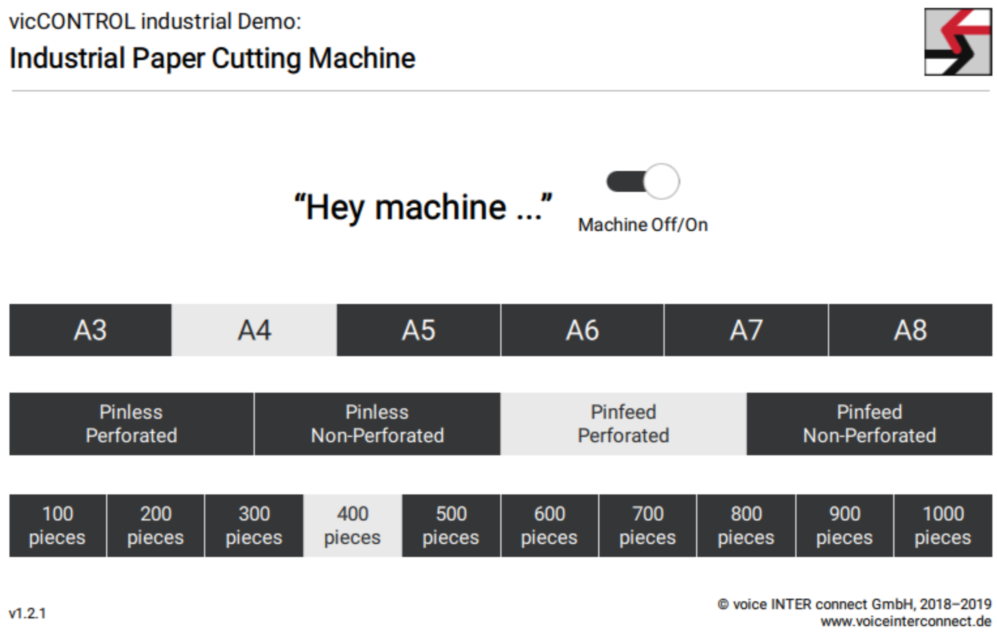
To implement the natural language understanding the open source solution Rasa NLU was used [9]. The intent classification is language independent and the slot value resolution was done automatically. Since the slot values were defined and included as word class tags during the statistical language model training, they were resolved by the SLM model itself.

The role of the dialogue manager is to verify the recognized intent – slot combinations, and depending on the configuration if a required slot is missing or the value cannot be verified, it will either eliciting a response from the speaker or end the dialogue while re-activating the wake-up-phrase recognizer. After classifying the intent and its slots, the semantic representation of the uttered and recognized command is sent over the MQTT broker to the control application. The control application implements the business logic, domain knowledge and the interface to external devices.

4 Industrial Voice Control Demonstrator

The industrial voice control demonstrator is built upon vicCONTROL industrial - embedded Voice Control for ARM Platforms [10]. The device provides hands-free and natural language interaction with arbitrary external systems via standard IoT or proprietary protocols. It does not require internet access nor remote speech services, the speech and audio feedback processing is performed on the embedded system, ensuring good accuracy, low latency and privacy by design. The following features are demonstrated: an SDK for embedded platforms, convenient API for using voice input and output functions via MQTT protocol, web-based designer toolkit vicSDC with application examples for the development of a customer-specific voice control system, industry-proven quality in 30 languages and dialects, wake-up word function and semantic evaluation for the recognized speech utterances by machine learning created NLU component.

The hardware of the demonstrator is based on the “phyBOARD-Mira” [11] and the NXP ARM Cortex-A9 i.MX6 series application processors. The board offers standard interfaces that include Ethernet, USB, RS-232/RS-485, CAN, a MicroSD card slot, and miniPCIe. It supports also HDMI, LVDS, parallel, resistive and capacitive touch, as well as other human-machine interaction options.



vicCONTROL industrial Demo:

Industrial Paper Cutting Machine

“Hey machine ...”

v1.2.1

© voice INTER connect GmbH, 2018–2019
www.voiceinterconnect.de

Figure 3 – Screenshot of the Industrial Paper Cutting Machine example application

Figure 3 presents the user interface (GUI) of one of the sample speech applications which is provided along with the target platform to demonstrate the simplicity of designing and expanding NLU speech interfaces for embedded applications. It simulates setting up the program of a paper cutting machine along with several input parameters in US English.

The speaker could switch on or off the machine and set the cutting parameters in a freely formulated manner with arbitrary slots in arbitrary order. The following intents with corresponding slots are defined: “set state” with slot “power” and “set-cutting-parameters” with slots: format (A3-A8), format (pin-less perforated, pin-less non-perforated, pin-feed perforated and pin-feed non-perforated) and quantity (100-1000 pieces).

The application can be easily expanded to other languages while preserving the same functionality and performance. The GUI is developed using the QT Framework and its behaviour is synchronized with the speech user interface.

5 Conclusions

This paper presents a workflow and a demonstrator for rapid design and deployment of multilingual embedded speech applications. The dialogues are modelled in a designer toolkit and the resources are deployed on a target embedded system.

The demonstrator presents hands-free and natural language interactions with arbitrary external systems via the standard Internet of Things protocols. It does not require internet access, the speech and audio feedback processing is performed offline, ensuring privacy by design. The presented approach allows rapid prototyping of multilingual dialogues, as a part of an iterative process of assessing and improving the quality of interaction and the user experience.

6 References

[1] STERLING, GREG. "*Google says 20 percent of mobile queries are voice searches. Voice search growing as virtual assistant market heats up.*" 18.05.2016, URL: https://searchengineland.com/google-reveals-20-percent-queries-voice-queries-249917, [retrieved on Oct 20, 2016]

[2] TRACTICA (2016). T*he virtual digital assistant market will reach $15.8 billion world-wide by 2021*, URL: https://www.tractica.com/newsroom/press-re-leases/the-virtual-digital-assistant-market-will-reach-15-8-billion-worldwide-by-2021/ (retrieved on Jan 6, 2019).

[3] RUPPERT, T., JASKÓ, S., HOLCZINGER, T. AND ABONYI, J., 2018. *Enabling Technologies for Operator 4.0: A Survey. Applied Sciences*, 8(9), p.1650.

[4] GRAY ST. *Always On: Privacy Implications of Microphone-Enabled Devices*. In: *Future of privacy forum,* 2016 Apr 9.

[5] ROMERO, D., STAHRE, J., WUEST, T., NORAN, O., BERNUS, P., FAST-BERGLUND, Å. AND GORECKY, D., 2016, October. *Towards an operator 4.0 typology: a human-centric perspective on the fourth industrial revolution technologies*. In: *INTERNATIONAL CONFERENCE ON COMPUTERS & INDUSTRIAL ENGINEERING (CIE46) (pp. 1-11).*

[6] R. JONSON, *Multilingual NLP Methods for Multilingual Dialogue Systems*, Dec. 2002, [retrieved on Dec 20, 2016], http://stp.lingfil.uu.se/~nivre/gslt/RebeccaNLP.pdf

[7] I. WENDLER, A. JATHO, I. KRALJEVSKI, M. WENZEL, *Nutzerzentrierter Entwurf von Multimodalen Bedien-konzepten*, 28. Konferenz Elektronische Sprach-signalverarbeitung 2017, Universität des Saarlandes, Saarbrücken, 15.–17. März 2017

[8] I. KRALJEVSKI, M. FISCHER, A. GJORESKI, D, HIRSCHFELD, *Development of a Natural Language Speech Dialogue System for an AR-based, Adaptive Mobility Agent*, 29. Conference on Electronic Speech Signal Processing 2018, Ulm University, March 7–9, 2018

[9] BOCKLISCH, T., FAULKNER, J., PAWLOWSKI, N., & NICHOL, A. (2017). *Rasa: Open Source Language Understanding and Dialogue Management.* CoRR, abs/1712.05181.

[10] Embedded Voice Control for ARM Platforms, vicCONTROL industrial. URL: https://www.voiceinterconnect.de/de/viccontrol_industrial [retrieved on Jan 20, 2019]

[11] phyBOARD®-Mira, ARM Cortex^{TM}-A9, URL: https://www.phytec.de/produkt/single-board-computer/phyboard-mira [retrieved on Jan 20, 2019]