Corpus generation for voice command in smart home and the effect of speech synthesis on End-to-End SLU

Thierry Desot, François Portet, Michel Vacher

Corpus generation for voice command in smart home and the effect of speech synthesis on End-to-End SLU

Thierry Desot, François Portet, Michel Vacher
Univ. Grenoble Alpes, CNRS, Grenoble INP, LIG
38000 Grenoble, France
Thierry.Desot@univ-grenoble-alpes.fr, {Francois.Portet,Michel.Vacher}@imag.fr

Abstract

Massive amounts of annotated data greatly contributed to the advance of the machine learning field. However such *large* data sets are often unavailable for novel tasks performed in realistic environments such as smart homes. In this domain, semantically annotated large voice command corpora for Spoken Language Understanding (SLU) are *scarce*, especially for non-English languages. We present the automatic generation process of a *synthetic* semantically-annotated corpus of French commands for smart-home to train pipeline and End-to-End (E2E) SLU models. SLU is typically performed through Automatic Speech Recognition (ASR) and Natural Language Understanding (NLU) in a pipeline. Since errors at the ASR stage reduce the NLU performance, an alternative approach is End-to-End (E2E) SLU to jointly perform ASR and NLU. To that end, the artificial corpus was fed to a text-to-speech (TTS) system to generate synthetic speech data. All models were evaluated on voice commands acquired in a real smart home. We show that artificial data can be combined with real data within the same training set or used as a stand-alone training corpus. The synthetic speech quality was assessed by comparing it to real data using dynamic time warping (DTW).

Keywords: Spoken language understanding, automatic speech recognition, natural language understanding, corpora and language resources, ambient intelligence, voice-user interface, text-to-speech, dynamic time warping

1. Introduction

Smart homes with integrated voice-user interfaces (VUI) can provide in-home assistance to older adults (Peetoom et al., 2015), allowing them to retain autonomy and providing swift intervention in emergency situations through distant interaction (Vacher et al., 2015). However these systems include multiple modules, such as a Spoken Language Understanding (SLU) module that must be able to extract the *intent* of the user from the voice command and its *named entities*. The intent reflects the intention of the speaker whereas entities and relations are called slots and represent the pieces of information that are relevant for the given task (Tur and De Mori, 2011). SLU systems typically consist of a pipeline of automatic speech recognition (ASR) and natural language understanding (NLU) modules. The NLU module takes as input a transcript of the voice command provided by the ASR module and extracts its meaning in a form that can be processed by a Decision Making module. The NLU model is trained on clean transcriptions whereas erroneous ASR transcriptions reduce the SLU performances. Different from pipeline SLU, the E2E SLU approach combines ASR and NLU in one model and avoids cumulative ASR and NLU errors (Ghannay et al., 2018). In (Qian et al., 2017) it is demonstrated that E2E SLU does not necessarily require an initial step of ASR if enough semantically annotated data is available.

SLU systems tend to impose a strict command syntax although senior adults interacting with smart environments are inclined to deviate from the imposed grammar of the commands (Takahashi et al., 2003; Möller et al., 2008; Vacher et al., 2015). Such systems are not flexible enough, which creates the need for a data driven SLU system, rather than a rule-based system. Unfortunately, large domain specific data sets are often not available, especially for languages other than English. For the French language, the closest data sets are either voice based but without voice commands (Fleury et al., 2013) or designed for other tasks (Chahuara et al., 2016). To deal with this data scarcity for the French language, we applied Natural Language Generation (NLG) to generate an artificial textual corpus, automatically labeled with named entity and intent classes. Similar to (Lugosch et al., 2019) using text-to-speech (TTS), artificial speech data was generated based on the synthetic corpus. Both data sets were used for training SLU models and evaluated on voice commands acquired in a real smart home with several speakers. (Desot et al., 2018; Mishakova et al., 2019; Desot et al., 2019b; Desot et al., 2019a).

The contributions of this paper are: 1) The generation of the first French semantically annotated synthetic corpus combined with artificial speech data for voice command in smart home. 2) The artificial data can be combined with realistic data in the same training set as shown in section 5. It can also be used as stand-alone training data, tested with realistic data. This paper gives an overview of the few available French corpora in the smart home domain in section 2. We present the *real* data test set that was recorded in a smart home in section 3. The generation and validation method of the *artificial* corpus and speech is outlined in section 4. followed by experiments, evaluation of the corpora in sections 5., 6. and a conclusion.

2. Comparable corpora

Due to an increasing interest in smart homes, speech corpora in this domain were recorded, especially for the English language. The CHiME-1 and CHiME-2 home automation corpora are recordings of 34 speakers, uttering 500 6-word commands based on a fixed grammar (Barker et al., 2017). The CHiME-5 corpus (Barker et al., 2018) is recorded in a *dinner party* context with 20 separate dinner sessions with two hosts and two guest participants.

Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020) , pages 6397–6406
Marseille, 11–16 May 2020

Each party lasts about two hours and was as natural as possible. The DIRHA English corpora (Ravanelli et al., 2015) include eleven hours of read and spontaneous home automation commands, keywords, phonetically-rich sentences and conversational speech from twelve native UK and US speakers. 50% of the data is based on simulations and the other half of the data on real recordings in several rooms. Similar to our generated artificial corpus, the utterances contain *keywords* to activate the home system.

For other languages than English, we mention the DICIT corpus (Brutti et al., 2008) recorded with four participants in a scenario of a distant-talking interface for interactive control of a TV. They were acting like a family while uttering phonetically rich sentences. This was combined with spontaneous interaction with the system. For the wizard-of-Oz experiments six hours of speech was recorded in English, German and Italian. The ITAAL Italian speech corpus (Principi et al., 2013) contains about 20 hours of recordings of 20 native Italian speakers at home uttering home automation commands and distress calls in normal and shouted conditions.

A Few French corpora in the smart home domain are available. However these are *not* or *partially* usable to train SLU models. The HIS corpus (Fleury et al., 2010) was one of the first including speech and everyday life recordings in a Health Smart Home. Different from the above mentioned corpora, it provides home automation sensor traces. 1886 individual sounds and 669 sentences were collected from 15 participants performing activities in a domestic context with a maximum duration of 1 h 35 minutes per experiment. The ANODIN-DETRESSE (AD) corpus was recorded in the context of emergency call detection in short phone calls by 21 older adults in a domestic context. The corpus includes 2646 utterances, annotated for a total duration of only 38 minutes (Vacher et al., 2008). Furthermore the utterances are read and do not contain any spontaneous speech. They are very short and thus lack syntactic and lexical variation. A similar corpus is the CIRDO corpus. This data set was recorded in realistic conditions in the SmartHome DOMUS[1], fully equipped with microphones and home automation sensors. 17 persons in the 40-60 age range performed scenarios including falls on a carpet and calls for help in the context of an audio/video emergency detection system. On average the acquisition duration was 2 hours and 30 minutes per person (Vacher et al., 2016). However the AD and CIRDO corpora do not cover more than one intent class. Another corpus with recordings of older adults in a domestic context is the ERES38 corpus (Entretiens RESidences 38: *Entretiens* means interviews). Contrary to the AD corpus it is a collection of annotated spontaneous speech (Aman et al., 2013). It was acquired from 22 senior adults between 68 and 98 years old. The total corpus includes 48 minutes of read speech and 17 hours of spontaneous speech.

We used the SWEET-HOME corpus *real* data that was collected in the smart home DOMUS, equipped with microphones for speech recording, sensors for providing information on the user's localization and activity (Vacher et al., 2014). It was recorded by participants enacting activities of daily living in a smart home equipped with home automation sensors and actuators. The recorded speech was mainly composed of voice commands. However it was collected with only single user settings with a set of commands respecting a strict grammar and is not sufficient to cover a large set of intents with a lot of syntactic and lexical variation. Characteristics of the SWEET-HOME corpus are included in Table 2.

We finally mention the VoiceHome-2 corpus for multichannel speech processing in real homes (Bertin et al., 2019). Overall, it contains 120 clean utterances, 360 reverberated utterances of distant-microphone spontaneous speech without noise and 1080 reverberated utterances with noise from 12 different native French speakers for a total duration of about 4 hours. All utterances are fully annotated with transcriptions and location for 12 real rooms. It does not contain any home automation sensor traces and is also not semantically annotated.

None of the above mentioned French corpora cover the intents and named entities defined by the Amiqual4Home smart home context (section 3.). On top of that combining those corpora does still not result in a sufficient amount of data, to train SLU models. For that reason we generated artificial data automatically labeled with slot and intent labels, defined by a smart home context.

3. VocADom@A4H corpus

Our corpus generator is easily adaptable to a modified smart home context and its target users. Its semantics are similar to the semi-automatically annotated slot labels and intent classes in the VocADom@A4H corpus (Portet et al., 2019; Desot et al., 2018). It includes about twelve hours of speech data and was acquired in realistic conditions in the Amiqual4Home smart home[2]. This two-storey 87m^2 smart home is equipped with home automation systems, multimedia devices, and microphone arrays. About 150 sensors and actuators were set in the house to acquire speech, to control lights, to set the heating etc. Eleven participants uttered voice commands while performing activities of daily living for about one hour in the kitchen, living room, bedroom and bathroom. Out-of-sight experimenters reacted to participants' voice commands following a wizard-of-Oz strategy to add naturalness to the corpus. To collect a corpus with spontaneous speech with lexical and syntactic variation, three recording phases were defined. Phase 1: Graphical based instruction to elicite spontaneous voice commands; phase 2: Two-inhabitant scenario between the dweller of the smart home and a visiting friend. Both utter spontaneous voice commands without grammar restrictions while interacting with the smart home; phase 3: Voice commands are recorded with background noise (vacuum cleaner, radio, tv etc.). In each voice command a keyword is used to activate the Smart Home. The resulting speech data was semi-automatically transcribed, annotated with intent classes and slot labels and resulted in 6,747 utterances (*complete(3)* in Table 1). It consists of voice commands (*intents(1)*), and other utterances than voice commands (*none intents(2)*).

[1] https://domus.liglab.fr

[2] https://amiqual4home.inria.fr

This *realistic* corpus is the held out test set used for all our SLU experiments as outlined in the next sections in order to assess the quality of the artificial corpus.

4. Artificial corpus generation

Syntactic variability and underspecification of commands make NLU development a challenging task. For the command *"raise the blinds"*, the NLU must identify the correct blinds in the home, based on the user's current location and activity. The same intent must be extracted from a more syntactically complex utterance such as *"can you raise the blinds"*. Similarly, *"a bit more"* following the command *"raise the blinds a bit"* must be *inferred* to be a request to repeat the previous action. The scope of our artificial corpus and SLU experiments, are commands without linguistic context and with one intent per utterance. We focused on the issue of syntactic and linguistic variability as occurring in the VocADom@A4H test corpus and had at the same time to tackle the linguistic distance between this realistic test set and the artificially generated training corpus.

4.1. Aligned and unaligned NLU transcriptions

For evaluation of the synthetic corpus, State-of-the-art NLU CRF models (Jeong and Lee, 2008) and also DNN-based models (Mesnil et al., 2015; Liu and Lane, 2016) were used. These models approach the NLU problem as a *sequence labeling task*. This means that the artificial training data must be *aligned* to associate each word to a slot label as in the *IOB* NE labeling scheme (inside, outside, beginning). The B-prefix before a tag indicates that the tag is the beginning of a NE and an I-prefix is used for a tag inside a NE. An O tag represents a token outside a NE. Using a *sequence generation task* with *unaligned* data, the model should learn to associate several words to one slot label without aligned data. For generation of the *aligned* and *unaligned* artificial corpus, standard expert-based NLG was chosen (Gatt and Krahmer, 2018) that can be controlled more easily as compared to a constrained RNN language model for data augmentation (Hou et al., 2018). The core of our corpus generator is the open source NLTK python library feature-based context free grammar (FCFG) (Bird et al., 2009), allowing for sentence generation, and for features (i.e. slot information) to be attached to the final output sentences.

The grammar defines intents as a composition of their possible constituents, with constraints on generation. For example, the *generative grammar rule* in table 3 defines the slots of the intent `set_device` and can generate the command *"open the window in the kitchen"*. `Slot_action` has the feature ACTION whereas `Slot_device` has the feature ALLOWABLE_ACTION. Both those features are set to the same variable value `?s` which makes sure we only generate phrases with an action that is applicable to a particular device. Subsequent rules, decomposing the constituents of the intent, contain other linguistic features such as gender and number agreement. Furthermore, domain constraints are defined for object location in the smart-home. To avoid the production of nonsensical utterances such as *"turn on the dishwasher in the bedroom"*, unification of features was applied. It is the process by which different symbols in rules are matched based on their features. If the rule defining for instance *"dishwasher"* device has a feature "location=[room="kitchen"]", the grammar must unify this feature with the same feature attributed to a room, in order to generate a sentence as for instance *"turn on the dishwasher in the kitchen"*.

Syntactical variation was also part of the grammar design, such as the French interrogative constructions with the particle (*"est-ce que"*) (Table 3 includes an example). Similar to the test set (section 3.), each voice command includes a keyword to activate the Smart Home. Maximizing all combinations of semantic labels that result in meaningful utterances, the grammar generates about 77,000 phrases, each annotated with an intent and slots for training purpose (Artif. in Table 2). An overview of intents is presented in Table 4. Slot labels are divided into eight *basic* categories: the `action` to perform, the `device` to act on, the `location` of the device or action, the `person` or `organization` to be contacted, a device `component`, a device `setting` and the `property` of a location, device, or world property. Together with variation on these categories, 17 slot labels are defined.

The artificial data is generated in three *aligned* formats to train NLU models for Rasa-NLU, Tri-CRF and Att-RNN using the *IOB* NE labeling scheme (Desot et al., 2018) and two *unaligned* formats for pipeline and E2E SLU (Desot et al., 2019b; Desot et al., 2019a). We give an example in *json* format for the utterance *can you close the blind*:

```
"vocadom tu peux fermer le store"
"intent": "set_device"
  {
  "start": 16,
  "end": 22,
  "entity": "action",
  "value": "close",
  "text": "fermer",
  },
  {
  "start": 23,
  "end": 31,
  "entity": "device",
  "value": "blind",
  "text": "le store",
  }
  ]
```

Table 5 (Tri-CRF) shows a sentence from the corpus version using the *IOB* NE labeling scheme format with a window of two words preceding and following the target word associated with a slot label for training the Tri-CRF model from (Jeong and Lee, 2008; Jeong and Lee, 2009). Table 5 (Att-RNN) shows the aligned corpus with the *IOB* NE labeling scheme for training the Att-RNN model from (Liu and Lane, 2016). There is a one-to-one mapping between source words and target labels: the French definite article 'le' is mapped to the target slot label 'B-person-occupation'.

Aligned approaches are less efficient for pipeline and E2E SLU with input data consisting of spontaneous speech with disfluencies. These frequently cause ASR deletion and in-

Table 1: SLU test data

VocADom@A4H	utterances	words	intents	slot labels	slot values
intents(1)	2612	430	7	14	60
none intents(2)	4135	1326	1	-	-
complete(3)	6747	1462	8	14	60

Table 2: Comparison of SLU training and test data (OOV = test set words not seen in training data)

training set	utterances	words	intents	slot labels	slot values	perplex.	OOV	VocADom@A4H
Artif.	77,481	187	7	17	69	124.41	307	intents(1)
Sweet-Home	1412	480	6	7	28	49.33	343	intents(1)
Eslo2	161,699	29,149	1	-	-	151.90	211	none intents(2)
Artif.+Sweet-Home+Eslo2	240,592	30,821	8	17	69	372.06	235	complete(3)

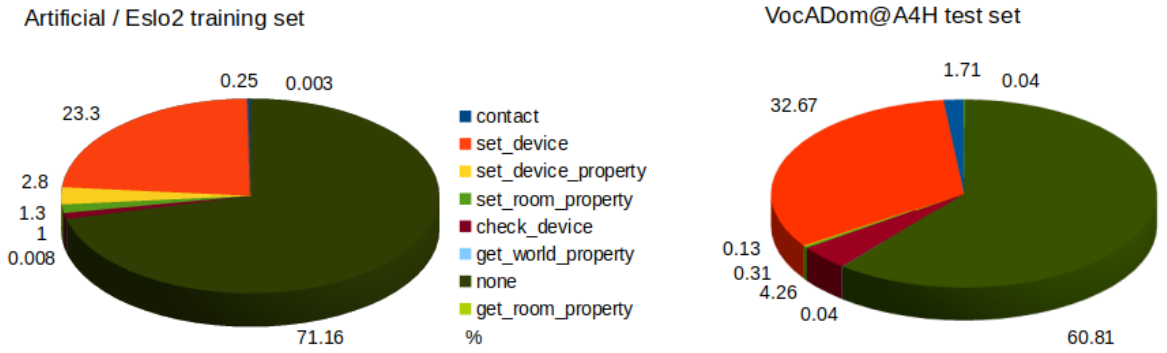

Figure 1: Intents in artificial/real training set and VocADom@A4H test set

sertion errors. Therefore an artificial data version was generated without alignment between source and target series of labels and intent classes for a *sequence generation* approach. Hence slot labels can be inferred from imperfect ASR transcriptions. The resulting unaligned corpus was the training data for an NLU seq2seq attention-based model[3] (Desot et al., 2019a; Desot et al., 2019b) approach. Table 5 includes a training example (Seq2seq). In this format, the intent is included (in square brackets) into the sequence of slot labels as first element. We assumed that the intent in initial position will improve the prediction of the following slots since these tend to depend on the intent. Slot-labels (in square brackets) are separated from slot-values so that models can learn them separately (Mishakova et al., 2019).

4.2. Transcriptions enriched with symbolic slot and intent labels

For the E2E SLU approach the artificial corpus transcriptions are enriched with intent class and slot label symbols (Desot et al., 2019b; Desot et al., 2019a). A similar approach was applied in (Ghannay et al., 2018). Transcriptions symbolically enriched with named entity labels were used to train a model with the Baidu Deep Speech ASR system (Hannun et al., 2014). Our approach is also inspired by (Serdyuk et al., 2018) where intents were directly inferred from audio MFCC features training a seq2seq model on clean and noisy speech data. Different from these approaches our transcriptions are enriched with *both* intent and slot label symbols. The symbolic labels per intent class are,

`set_device intent`	`@`
`set_device_property`	`_`
`set_room_property`	`&`
`check_device`	`#`
`get_world_property`	`]`
`get_room_property`	`{`
`contact`	`[`

An example of an enriched transcription is included in the last two rows of Table 5. As an E2E approach extracts slot labels and intent classes directly from speech, the second part of the E2E artificial corpus, is an artificial speech data base. To that end synthetic speech was generated for the 77k artificial corpus sentences, using the open source ubuntu SVOX [4] female French voice [5]. An E2E model was trained on the synthetic speech source data and the symbolically enriched artificial target data transcriptions using ESPnet (Watanabe et al., 2018).

4.3. Acoustic validity of the artificial speech data

As the artificial speech data is a part of our corpus, we assessed its quality by comparing the artificial speech to the VocADom@A4H real speech data, by calculating the acoustic distances between the two data sets. We generated TTS for the 6747 utterances of the VocaDom@A4H test set and calculated the acoustic distance between the *real speech* utterances and the resulting *artificial speech* utterances using dynamic time warping (DTW). The DTW

[3] https://gricad-gitlab.univ-grenoble-alpes.fr/getalp/seq2seqpytorch

[4] https://launchpad.net/ubuntu/+source/svox

[5] https://doc.ubuntu-fr.org/svoxpico

Table 3: Syntactic variation with annotation in the artificial corpus and grammar rule

Sentence (French)	**English translation**
Ouvre la fenêtre dans la cuisine	*Open the window in the kitchen*
Syntactic variation	
Est-ce que tu peux ouvrir la fenêtre dans la cuisine?	*Can you open the window in the kitchen?*
Annotation	
`SET_DEVICE(ACTION=open="open",DEVICE=window="window",LOCATION=room="kitchen")`	
Generative grammar rule	
`Intent_set_device[ACTION=?s,Location=?l, Device=?d] → Slot_action[ACTION=?s, :ACTION_TYPES={}, AGR=?a] Slot_device[ALLOWABLE_ACTION=?s, Location=?l, Device=?d, ARTTYPE=def]`	

Table 4: Artificial corpus (Artif.) and VocADom@A4H (Real.): Examples and Frequency of intents

Intent	**Example (French)**	**English translation**	**Frequency**	
			Artif.	**Real.**
`Contact`	*Appelle un médecin*	*Call a doctor*	567	114
`Set_device`	*Ouvre la fenêtre*	*Open the window*	63,288	2178
`Set_device_property`	*Diminue le volume de la télé*	*Decrease the TV volume*	7290	9
`Set_room_property`	*Diminue la température*	*Decrease the temperature*	3564	21
`Check_device`	*Est-ce que la fenêtre est ouverte?*	*Is the window open?*	2754	284
`Get_room_property`	*Quelle est la température?*	*What's the temperature?*	9	3
`Get_world_property`	*Quelle heure est-il?*	*What's the time?*	9	3
`None`	*La fenêtre est ouverte*	*The window is open*	-	4135

distance measure is a technique that has been introduced in speech recognition a few decades ago by Sakoe *et al.* (Sakoe and Chiba, 1978) and is still in use (Dhingra et al., 2013; Su et al., 2019). Since time alignment of different utterances is a core problem for distance measurement of speech sequences, DTW measures the similarity between two time series which may vary or *warp* in time. The optimal alignment is found for a time series that is warped non-linearly by stretching or shrinking it along its time axis. This similarity is measured with the minimum edit distance. Thus, two identical time series will have a DTW distance of zero (Muda et al., 2010; Sakoe and Chiba, 1978). Time series Q and C of length n and m respectively, $Q = q_1, q_2, ..., q_i, q_n$ and $C = c_1, c_2, ..., c_j, c_m$, are aligned in an n-by-m matrix, using DTW, where n is the number of frames in the first and m the number of frames in the second signal. The (*i*th, *j*th) element of the matrix contains the distance $d(q_i, c_j)$ between the two points q_i and c_j. Each matrix element (i, j) corresponds to the alignment between the points q_i and c_j. The accumulated distance is measured by:

$$D(i,j) = min[D(i-1,j-1), D(i-1,j), D(i,j-1)] + d(i,j) \quad (1)$$

The absolute distance between the values of the two sequences is calculated using the Euclidean distance:

$$d(q_i, c_j) = (q_i - c_j)^2 \quad (2)$$

This is shown in Figure 2 for a sample of real speech (VocADom@A4H) with 20 mfcc's over 168 frames, and a TTS sample (20 mfcc's over 179 frames) for the French sentence *"chanticou arrêtez les stores de la salle de bains"*, *chanticou stop (opening or closing) the blinds in the bathroom.* The blue line in Figure 2 shows the optimal warping path which minimizes the sum of the DTW distance between the artificial and the real speech signal. The darker regions show a higher cost and distance. Using the python librosa library, DTW was calculated on 20 mfcc features from the original TTS and real speech 16 kHz wav files.

Table 6 includes the DTW between (female) TTS and real samples (*TTS vs. real*), and inter speaker DTW for real samples for identical sentences (*Inter-real*). The distances were normalized by dividing the total distance by the length of the longest time series (*Long norm.*), by the length of the shortest time series (*Short norm.*), and by the length of the optimal warping path (*Opt. norm.*) (Table 6) (Ratanamahatana and Keogh, 2004). In Table 6 the average distance and standard deviation over all compared samples are mentioned. For a comparison between (female) TTS and real speech, DTW was calculated between the 6747 TTS and all corresponding real speech samples (*all*). Table 6 also includes a comparison between all real speech male samples and the corresponding number (between brackets) of TTS samples (*male*). The third row includes a comparison between all real speech female samples and the corresponding TTS samples (*female*).

For real speech inter speaker comparison (last row), we calculated DTW for sentences uttered by *all* speakers (4 common sentences, *common*). Thus, for each speaker we compared the time series with all time series of the other speakers uttering the same sentence once, or more than once, and calculated DTW for each pair of time series. Table 6 shows that in general inter speaker distances for real speech are significantly smaller than between TTS and real speech. As the artificial speech is generated for a French *female* voice, distances between TTS and real female samples are smaller as compared to distances between real male and TTS sam-

Table 5: Aligned, unaligned artificial corpus format and symbolically enriched transcriptions

Aligned

Tri-CRF ("vocadom call a doctor")

```
(Source) vocadom appelle médecin
(Target)
O vocadom /                                  -1=<s>      +1=appelle +2=médecin
action-B appelle /            -2=<s>         -1=vocadom  +1=médecin +2=</s>
person-occupation-B médecin / -2=vocadom     -1=appelle  +1=</s>
CONTACT
```

Att-RNN ("vocadom call the doctor")

```
(Source) vocadom appelle  le                    médecin
(Target) O       B-action B-person-occupation I-person-occupation
CONTACT
```

Unaligned

Seq2seq ("vocadom close the door")

```
(Source) vocadom ferme la porte
(Target) intent[set_device], action[close], device[door]
```

Symbolically enriched

E2E (ESPnet) ("vocadom switch on the light")

```
(Source + Target labels injected)
@ VocADom                  ^allume^          }la lumière} @
```

SET_DEVICE intent class symbol @/ Action slot symbol ^ / Device slot symbol }

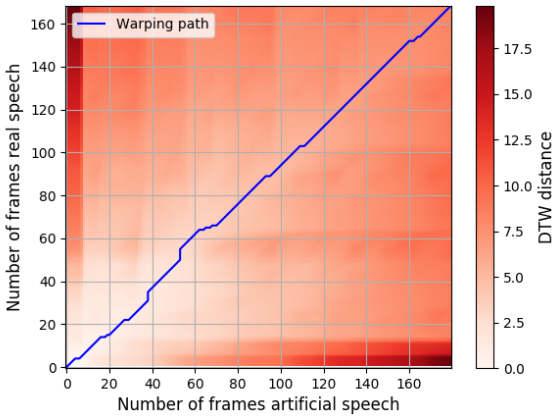

Figure 2: Warping path and DTW distance

ples. Experiments in section 5. show the impact of the distance between *artificial* training and *real* test data on the *symbolic* and on the *acoustic level*. On top of that we outline how we dealt with the bottleneck of combining synthetic and real speech in the training data.

Table 6: DTW TTS vs. real speech, mean-standard deviation

DTW	**Long norm.**	**Short norm.**	**Opt norm.**
TTS vs. real:			
all (6747)	$5.58^{\pm 4.42}$	$7.20^{\pm 8.32}$	$4.71^{\pm 3.61}$
male (4372)	$5.72^{\pm 4.02}$	$7.31^{\pm 7.86}$	$4.83^{\pm 3.28}$
female (2375)	$\mathbf{5.32^{\pm 5.04}}$	$\mathbf{7.01^{\pm 9.07}}$	$\mathbf{4.51^{\pm 4.12}}$
Inter-real:			
common (2806)	$1.85^{\pm 2.22}$	$4.16^{\pm 9.05}$	$1.67^{\pm 2.07}$

5. Evaluation experiments and results

For evaluation of the artificial corpus and speech data, we examined performances for aligned and unaligned NLU, a pipeline and E2E SLU approach, *with* and *without* real data in the training set. Performances are included in Table 7. The data sets in the first column are specified in Figure 3.

5.1. NLU

The training data for all aligned approaches is the *artificial* corpus *only*, using the VocADom@A4H *real* test set. The Att-RNN models ((3) and (4) in Table 7 and Figure 3), based on aligned data, outperform the Rasa-NLU (1) and Tri-CRF (2) models and show the feasibility of using artificial NLU training data and realistic held-out test data, in spite of the linguistic distance between both data sets (Desot et al., 2018). The Seq2seq model (5), using the same amount of artificial training data but *unaligned*, augmented with 727 utterances of realistic domain specific SWEET-HOME data, is only slightly outperformed by the Att-RNN model for intent prediction. However it indicates that a model trained with unaligned data can be tested with ASR output transcriptions and be integrated in a pipeline SLU approach. The 3-gram SWEET-HOME corpus language model (LM) perplexity (*perplex.* in Table 2) on the test set sentences with voice command is 49.33 and significantly lower as compared to the artificial data LM perplexity on the VocADom@A4H test set. Integrating these real data in the training set partially contributes to boost the seq2seq model performances. As shown in Figure 1 and Table 1 (*none intents(2)*), `none` intents are the majority class in the VocADom@A4H test set. To model the `none` intent we increased the training data set with ESLO2 corpus (Table 2) utterances of conversational French speech (Serpollet et al., 2007). Similar to the VocADom@A4H and SWEET-HOME corpora, it contains frequent disfluen-

cies. Sentences which were unrelated to voice command intent were extracted (i.e. `none` intent) and manually filtered. Only out of domain utterances were kept for collecting `none` intent training data. Table 2 shows the lowest OOV of `none` intent test utterances (*none intents(2)*) as compared to the ESLO2 training set. The complete test set (*complete(3)*) and full training set for the seq2seq model (*Artif.+Sweet-Home+Eslo2*) are specified in the last row.

Table 7 (Seq2seq(6)) includes performances significantly lower than the other Att-RNN and Seq2seq models due to a strong tendency towards `none` intent prediction with ESLO2 data as part of the training set (Figure 3 Seq2seq(6)) (Desot et al., 2019a; Desot et al., 2019b).

Table 7: aligned and unaligned NLU performances (%) on VocADom@A4H

NLU Model +**Data set**	**Intent F1-score**	**Slot F1-score**
Aligned:		
Rasa-NLU(1)	76.57	**79.03**
Tri-CRF(2)	76.36	60.64
Att-RNN1(3)	91.30	66.09
Att-RNN2(4)	**96.70**	74.27
Unaligned:		
Seq2seq1(5)	**94.74**	51.06
Seq2seq2(6)	85.51	**65.49**

5.2. SLU

For the pipeline SLU approach, a large acoustic model (*Kaldi-Seq2seq-complete(7)* in Table 8 and Figure 3) was trained using 90% of the 472.65 hours of *Real data* in Figure 3, the other 10% being the development set. This data includes the corpora ESTER1 (Galliano et al., 2005) and 2 (Galliano et al., 2009), REPERE , ETAPE, BREF120 (Tan and Besacier, 2006), AD, SWEET-HOME and CIRDO (section 2.). The ASR transcriptions were generated using the hybrid HMM-DNN Kaldi tool using speaker adapted features from the Gaussian mixture model (GMM) (Povey et al., 2011). Its output transcriptions were fed to the seq2seq NLU module outlined in section 5.1. (Seq2seq(6) in Table 7 and Figure 3), (Desot et al., 2019a; Desot et al., 2019b). We report NLU performances on the VocADom@A4H test set using the concept error rate (CER) for slot labels. As the NLU problem is designed as a sequence generation task using unaligned data, the type of errors differs from a sequence labeling task with aligned data. Typical errors using aligned data are substitutions whereas with unaligned data, frequent deletions and insertions occur. In (Hahn et al., 2008) the CER is defined as the ratio of the sum of deleted, inserted and confused concepts w.r.t. a Levenshtein-alignment for a given reference concept string. We compared with a *small* pipeline model consisting of an acoustic model trained with ESPnet and an NLU seq2seq model both trained on 94.39% of *artificial* data (*ESPnet-Seq2seq-small(8)* in Table 8 and in Figure 3). ESPnet was used for the acoustic model (60.6% WER), as Kaldi performances dramatically decreased integrating artificial speech in the training data (WER >90%). The small model, with almost completely artificial data, shows slot label predictions similar to the large model with only real data.

For the E2E experiments, we used ESPnet default settings (Desot et al., 2019a) in order to train on speech data, with slots and intents symbolically injected in the transcriptions (section 4.2.). We also injected the intent and slot symbols into the clean transcriptions of the 553.9 hours of speech training data utterances. These were bootstrapped from the artificial data. In the sentence *"The light is switched off" (La lumière est éteinte)*, the slot label for *"The light"* is `device`. A model with real and artificial data (553.9h), a second model of only artificial speech (81.25h), and a third model of predominantly artificial speech (84.69) were trained (respectively *ESPnet-complete(9)*, *ESPnet-Artif-only(10)* and *ESPnet-small(11)* in Table 8 and Figure 3). It shows that E2E SLU with a small training set consisting of predominantly artificial domain-specific speech is feasible. Adding a small portion of real data to the training model significantly improves performances, indicating a too large distance between real test and artificial training data acoustic features.

Table 8: Pipeline and E2E SLU performances (% F1-score - Concept Error Rate) on VocADom@A4H

SLU Model +**Data set**	**(%) TTS in train**	**Intent F1-score**	**Slot CER**
Pipeline:			
Kaldi-Seq2seq-complete(7)	0.00	**84.21**	36.24
ESPnet-Seq2seq-small(8)	94.39	61.35	**35.62**
E2E:			
ESPnet-complete(9)	14.67	47.31	51.87
ESPnet-Artif-only(10)	100.00	35.94	56.00
ESPnet-small(11)	94.39	**75.73**	**26.17**

Table 9: E2E SLU performances (%) on male and female VocADom@A4H utterances

SLU Model +**Data set**	**(%) TTS in train**	**Intent F1-score**	**Slot CER**
E2E:			
ESPnet-small(11) all	94.39	75.73	26.17
ESPnet-small(11) female	94.39	**77.00**	**25.45**
ESPnet-small(11) male	94.39	75.11	28.70

Table 9 repeats best performances for the E2E SLU model with 94.39% of artificial speech in the training data (*ESPnet-small(11)* from Table 8), and includes performances for separate female and male VocADom@A4H real test data(female/male in the last two rows). Similar to smaller distances between (female) TTS and real *female* samples (section 4.3.), E2E SLU results are slightly better for the female test data than for the male test data

6. Discussion

The Att-RNN2 NLU (Att-RNN(4) in Table 7) outperforms the other aligned and unaligned models. It is competitive with the state of the art NLU showing that for an NLU task, the artificial corpus can be used as a stand alone corpus, in spite of the the lexical and syntactic distance be-

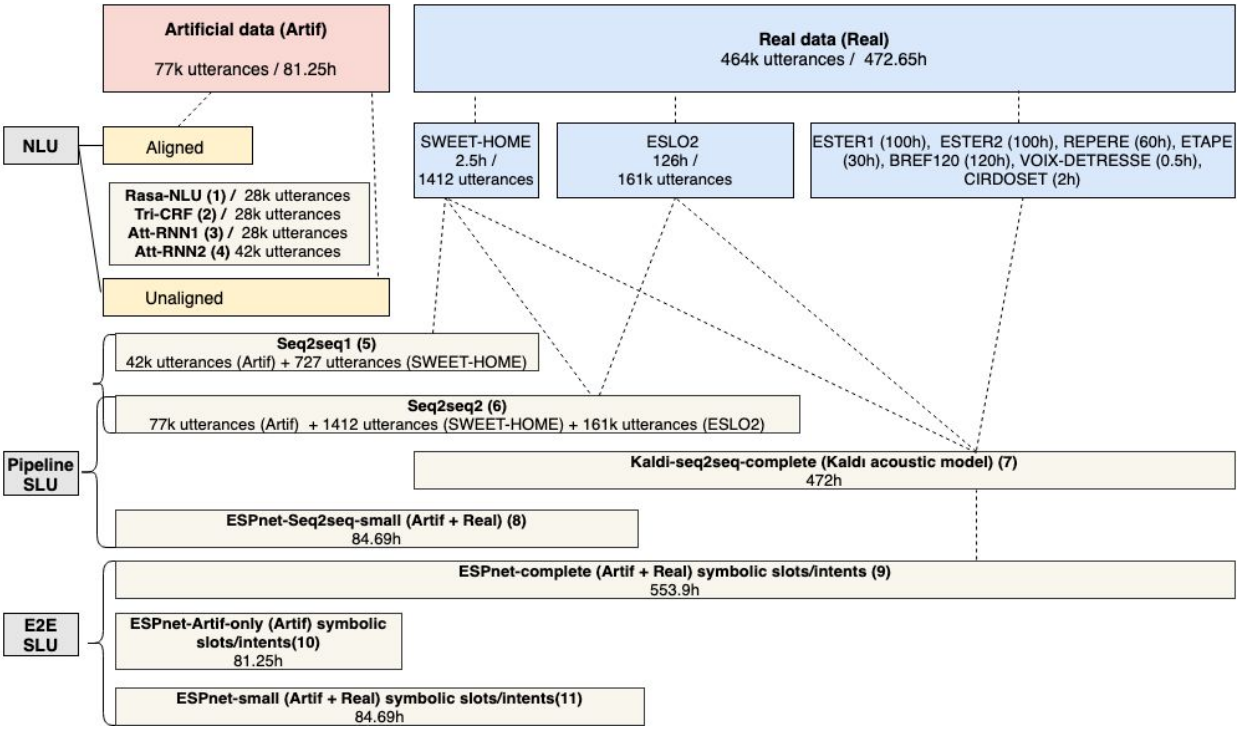

Figure 3: Training data sets and models overview

tween realistic test and artificial data. We used standard expert-based NLG (Gatt and Krahmer, 2018) in order to create the artificial corpus. However, for generating sentences with more lexical and syntactic variation, closer to the VocADom@A4H test data, *Generate Adversary Networks* (GAN) text generation techniques should be studied (Yu et al., 2017) where the *discriminator* tries to differentiate between real and fake data produced by the *generator*. For pipeline SLU, poor performances are shown in Table 8 (Pipeline, *ESPnet-seq2seq-small(8)*) with training data predominantly consisting of TTS (94.39%). This is contrary to E2E SLU performances with the same data size (E2E, *ESPnet-small(11)*), making better use of a reduced data set, predominantly being *artificial* data. For the Pipeline SLU model, optimal performances are only exhibited using an ASR module trained on a huge amount of *real* speech data (472 h) (Kaldi acoustic model (7) in Figure 3).

For E2E SLU Table 9 exhibits better performances for separate female VocADom@A4H real test data (*ESPnet-small(11) female* in the last row) as compared to only male test data. This is due to the (only) female artificial speech in the training data and is in line with DTW *acoustic* distances being larger between (female) artificial speech and male real speech as compared to female real speech (Table 6). Subsequently male TTS speech should be added to boost performances. On top of that real inter speaker DTW acoustic distances are significantly smaller than acoustic distances between TTS and real speech (Table 6). This indicates that TTS generated with other TTS voices might decrease this distance. Another possibility is to train a neural speech synthesis model such as *Tacotron* (Wang et al., 2017; Li et al., 2018) for SWEET-HOME real corpus data, and train with TTS for the synthetic corpus sentences. A *symbolic* analysis shows frequent E2E ASR errors for the keyword proper noun predictions (10% of the total ASR errors), partially due to TTS mispronunciations.

7. Conclusion

In this study we demonstrated the feasibility to train SLU models, with an artificial semantically labeled domain-specific corpus. We made our artificial corpus available to the community[6], including the data in different formats to train state-of-the-art aligned, unaligned NLU models, and the enriched transcriptions with symbolically inserted slot and intent labels, to train E2E SLU models. Corpus evaluation using the E2E SLU model exhibits promising results and shows the possibility of using artificial data as almost a stand alone training set. Different from pipeline SLU, an E2E model can be trained with small domain specific data sets, artificially generated. *Acoustic* distance analysis shows that the artificial (female) speech data should be augmented with male voice TTS. The training set can also be augmented with TTS data as output from neural speech synthesis models trained on real domain specific training data. NLG experiments using GANs, might be considered in order to decrease the *symbolic*, lexical and syntactic distance between artificial and real data. Future work to improve E2E SLU performances, includes multi-task and transfer learning.

8. Acknowledgements

This work is part of the VocADomproject founded by the French National Research Agency (Agence Nationale de la Recherche) / ANR-16-CE33-0006.

[6] https://gricad-gitlab.univ-grenoble-alpes.fr/Vocadom

9. Bibliographical References

Aman, F., Vacher, M., Rossato, S., and Portet, F. (2013). Speech recognition of aged voice in the aal context: Detection of distress sentences. In *2013 7th Conference on Speech Technology and Human-Computer Dialogue (SpeD)*, pages 1–8. IEEE.

Barker, J., Marxer, R., Vincent, E., and Watanabe, S. (2017). The CHiME challenges: Robust speech recognition in everyday environments. In *New Era for Robust Speech Recognition - Exploiting Deep Learning*, pages 327–344. Springer, November.

Barker, J., Watanabe, S., Vincent, E., and Trmal, J. (2018). The fifth'chime'speech separation and recognition challenge: Dataset, task and baselines. *arXiv preprint arXiv:1803.10609*.

Bertin, N., Camberlein, E., Lebarbenchon, R., Vincent, E., Sivasankaran, S., Illina, I., and Bimbot, F. (2019). Voicehome-2, an extended corpus for multichannel speech processing in real homes. *Speech Communication*, 106:68–78.

Bird, S., Klein, E., and Loper, E. (2009). *Natural Language Processing with Python*. O'Reilly Media.

Brutti, A., Cristoforetti, L., Kellermann, W., Marquardt, L., and Omologo, M. (2008). WOZ acoustic data collection for interactive TV. In *Proceedings of 6th Int. Conf. on Language Resources and Evaluation (LREC)*, pages 2330–2334.

Chahuara, P., Fleury, A., Portet, F., and Vacher, M. (2016). On-line Human Activity Recognition from Audio and Home Automation Sensors: comparison of sequential and non-sequential models in realistic Smart Homes. *Journal of ambient intelligence and smart environments*, 8(4):399–422.

Desot, T., Raimondo, S., Mishakova, A., Portet, F., and Vacher, M. (2018). Towards a French Smart-Home Voice Command Corpus: Design and NLU Experiments. In *21st International Conference on Text, Speech and Dialogue TSD 2018*, Brno, Czech Republic.

Desot, T., Portet, F., and Vacher, M. (2019a). Slu for voice command in smart home: comparison of pipeline and end-to-end approaches. In *IEEE Automatic Speech Recognition and Understanding Workshop (ASRU 2019)*, Sentosa, Singapore.

Desot, T., Portet, F., and Vacher, M. (2019b). Towards end-to-end spoken intent recognition in smart home. In *The 10th Conference on Speech Technology and Human-Computer Dialogue (SpeD 2019)*, Timisoara, Romania.

Dhingra, S. D., Nijhawan, G., and Pandit, P. (2013). Isolated speech recognition using mfcc and dtw. *International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering*, 2(8):4085–4092.

Fleury, A., Vacher, M., Portet, F., Chahuara, P., and Noury, N. (2010). A multimodal corpus recorded in a health smart home. In *Proceedings of LREC Workshop Multimodal Corpora and Evaluation*, pages 99–105, Malta.

Fleury, A., Vacher, M., Portet, F., Chahuara, P., and Noury, N. (2013). A french corpus of audio and multimodal interactions in a health smart home. *Journal on Multimodal User Interfaces*, 7(1):93–109.

Galliano, S., Geoffrois, E., Mostefa, D., Choukri, K., Bonastre, J.-F., and Gravier, G. (2005). The ester phase ii evaluation campaign for the rich transcription of French broadcast news. In *Ninth European Conference on Speech Communication and Technology*.

Galliano, S., Gravier, G., and Chaubard, L. (2009). The ester 2 evaluation campaign for the rich transcription of French radio broadcasts. In *Tenth Annual Conference of the International Speech Communication Association*.

Gatt, A. and Krahmer, E. (2018). Survey of the state of the art in natural language generation: Core tasks, applications and evaluation. *Journal of Artificial Intelligence Research*, 61(1).

Ghannay, S., Caubrière, A., Estève, Y., Camelin, N., Simonnet, E., Laurent, A., and Morin, E. (2018). End-to-end named entity and semantic concept extraction from speech. In *IEEE Spoken Language Technology Workshop*, Athens, Greece.

Hahn, S., Lehnen, P., Raymond, C., and Ney, H. (2008). A comparison of various methods for concept tagging for spoken language understanding. In *LREC 2008*.

Hannun, A., Case, C., Casper, J., Catanzaro, B., Diamos, G., Elsen, E., Prenger, R., Satheesh, S., Sengupta, S., Coates, A., et al. (2014). Deep speech: Scaling up end-to-end speech recognition. *arXiv preprint arXiv:1412.5567*.

Hou, Y., Liu, Y., Che, W., and Liu, T. (2018). Sequence-to-sequence data augmentation for dialogue language understanding. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1234–1245.

Jeong, M. and Lee, G. G. (2008). Triangular-chain conditional random fields. *IEEE Transactions on Audio, Speech, and Language Processing*, 16(7):1287–1302.

Jeong, M. and Lee, G. G. (2009). Multi-domain spoken language understanding with transfer learning. *Speech Communication*, 51(5):412–424.

Li, J., Gadde, R., Ginsburg, B., and Lavrukhin, V. (2018). Training neural speech recognition systems with synthetic speech augmentation. *arXiv preprint arXiv:1811.00707*.

Liu, B. and Lane, I. (2016). Attention-based recurrent neural network models for joint intent detection and slot filling. In *Proceedings of Interspeech 2016*, pages 685–689.

Lugosch, L., Meyer, B., Nowrouzezahrai, D., and Ravanelli, M. (2019). Using speech synthesis to train end-to-end spoken language understanding models. *arXiv preprint arXiv:1910.09463*.

Mesnil, G., Dauphin, Y., Yao, K., Bengio, Y., Deng, L., Hakkani-Tur, D., He, X., Heck, L., Tur, G., Yu, D., and others. (2015). Using recurrent neural networks for slot filling in spoken language understanding. *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, 23(3):530–539.

Mishakova, A., Portet, F., Desot, T., and Vacher, M. (2019). Learning Natural Language Understanding Systems from Unaligned Labels for Voice Command in

Smart Homes. In *The 1st International Workshop on Pervasive Computing and Spoken Dialogue Systems Technology (PerDial 2019)*, Kyoto, Japan.

Möller, S., Gödde, F., and Wolters, M. (2008). Corpus analysis of spoken smart-home interactions with older users. In *Proceedings of the 6th International Conference on Language Resources and Evaluation*.

Muda, L., Begam, M., and Elamvazuthi, I. (2010). Voice recognition algorithms using mel frequency cepstral coefficient (mfcc) and dynamic time warping (dtw) techniques. *arXiv preprint arXiv:1003.4083*.

Peetoom, K. K. B., Lexis, M. A. S., Joore, M., Dirksen, C. D., and De Witte, L. P. (2015). Literature review on monitoring technologies and their outcomes in independently living elderly people. *Disability and Rehabilitation. Assistive Technology*, 10(4):271–294.

Portet, F., Caffiau, S., Ringeval, F., Vacher, M., Bonnefond, N., Rossato, S., Lecouteux, B., and Desot, T. (2019). Context-Aware Voice-based Interaction in Smart Home -VocADom@A4H Corpus Collection and Empirical Assessment of its Usefulness. In *17th IEEE International Conference on Pervasive Intelligence and Computing (PICom 2019)*, Fukuoka, Japan.

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., et al. (2011). The Kaldi speech recognition toolkit. Technical report, IEEE Signal Processing Society.

Principi, E., Squartini, S., Piazza, F., Fuselli, D., and Bonifazi, M. (2013). A distributed system for recognizing home automation commands and distress calls in the italian language. In *Interspeech 2013*, pages 2049–2053.

Qian, Y., Ubale, R., Ramanaryanan, V., Lange, P., Suendermann-Oeft, D., Evanini, K., and Tsuprun, E. (2017). Exploring asr-free end-to-end modeling to improve spoken language understanding in a cloud-based dialog system. In *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 569–576. IEEE.

Ratanamahatana, C. A. and Keogh, E. (2004). Everything you know about dynamic time warping is wrong. In *Third workshop on mining temporal and sequential data*, volume 32. Citeseer.

Ravanelli, M., Cristoforetti, L., Gretter, R., Pellin, M., Sosi, A., and Omologo, M. (2015). The DIRHA-English corpus and related tasks for distant-speech recognition in domestic environments. In *Automatic Speech Recognition and Understanding (ASRU)*, pages 275–282.

Sakoe, H. and Chiba, S. (1978). Dynamic programming algorithm optimization for spoken word recognition. *IEEE transactions on acoustics, speech, and signal processing*, 26(1):43–49.

Serdyuk, D., Wang, Y., Fuegen, C., Kumar, A., Liu, B., and Bengio, Y. (2018). Towards end-to-end spoken language understanding. *arXiv preprint arXiv:1802.08395*.

Serpollet, N., Bergounioux, G., Chesneau, A., and Walter, R. (2007). A large reference corpus for spoken French: Eslo 1 and 2 and its variations. In *Proceedings from Corpus Linguistics Conference Series, University of Birmingham*.

Su, H., Dzodzo, B., Wu, X., Liu, X., and Meng, H. (2019). Unsupervised methods for audio classification from lecture discussion recordings. *Proc. Interspeech 2019*, pages 3347–3351.

Takahashi, S.-y., Morimoto, T., Maeda, S., and Tsuruta, N. (2003). Dialogue Experiment for Elderly People in Home Health Care System. In *Text, Speech and Dialogue*, pages 418–423, Brno, Czech Republic.

Tan, T.-P. and Besacier, L. (2006). A French non-native corpus for automatic speech recognition. In *Proceedings of the Language Resources and Evaluation Confere (LREC)*, volume 6, pages 1610–1613.

Tur, G. and De Mori, R. (2011). *Spoken Language Understanding Systems for Extracting Semantic Information from Speech*. Wiley.

Vacher, M., Fleury, A., Serignat, J.-F., Noury, N., and Glasson, H. (2008). Preliminary evaluation of speech/sound recognition for telemedicine application in a real environment. In *Interspeech'08*, pages 496–499, Brisbane, Australia.

Vacher, M., Lecouteux, B., Chahuara, P., Portet, F., Meillon, B., and Bonnefond, N. (2014). The Sweet-Home speech and multimodal corpus for home automation interaction. In *9th Language Resources and Evaluation Conference (LREC)*, pages 4499–4506, Reykjavik, Iceland.

Vacher, M., Caffiau, S., Portet, F., Meillon, B., Roux, C., Elias, E., Lecouteux, B., and Chahuara, P. (2015). Evaluation of a context-aware voice interface for ambient assisted living: qualitative user study vs. quantitative system evaluation. *ACM Transactions on Accessible Computing*, 7(2):5:1–5:36.

Vacher, M., Bouakaz, S., Chaumon, M.-E. B., Aman, F., Khan, R. A., Bekkadja, S., Portet, F., Guillou, E., Rossato, S., and Lecouteux, B. (2016). The cirdo corpus: comprehensive audio/video database of domestic falls of elderly people. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 1389–1396.

Wang, Y., Skerry-Ryan, R., Stanton, D., Wu, Y., Weiss, R. J., Jaitly, N., Yang, Z., Xiao, Y., Chen, Z., Bengio, S., et al. (2017). Tacotron: Towards end-to-end speech synthesis. *arXiv preprint arXiv:1703.10135*.

Watanabe, S., Hori, T., Karita, S., Hayashi, T., Nishitoba, J., Unno, Y., Soplin, N., Heymann, J., Wiesner, M., Chen, N., Renduchintala, A., and Ochiai, T. (2018). Espnet: End-to-end speech processing toolkit. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, pages 2207–2211.

Yu, L., Zhang, W., Wang, J., and Yu, Y. (2017). Seqgan: Sequence generative adversarial nets with policy gradient. In *Thirty-First AAAI Conference on Artificial Intelligence*.