ISCA Archive
http://www.isca-speech.org/archive

INTERSPEECH 2012
ISCA's 13th Annual Conference
Portland, OR, USA
September 9-13, 2012

Automatic word naming recognition for treatment and assessment of aphasia

Alberto Abad, Anna Pompili, Angela Costa, Isabel Trancoso

L^{2}F - Spoken Language Systems Lab, INESC-ID Lisboa, Portugal
alberto@l2f.inesc-id.pt

10.21437/Interspeech.2012-310

Abstract

VITHEA is an on-line platform designed to act as a "virtual therapist" for the treatment of Portuguese speaking aphasic patients. Concretely, the system integrates automatic speech recognition technology to provide word naming exercises to individuals with lost or reduced word naming ability. In this paper, we present the solution adopted for the word naming task, which is based on a keyword spotting approach with hybrid HMM/MLP speech recognizer. Furthermore, we explore a simple cross-validation method that makes use of the patients measured word naming ability to automatically adapt to their speech particularities. A corpus with word naming therapy sessions of aphasic Portuguese native speakers has been collected to test the utility of the approach for both global evaluation and treatment. In spite of the different patient characteristics and speech quality conditions of the collected data, encouraging results have been obtained.

Index Terms: speech disorders, aphasia, word naming

1. Introduction

The number of individuals that suffer cerebral vascular accidents has increased in the last decades in the EU, being estimated that a third of the patients present language deficiencies [1]. Aphasia is a particular type of language disorder that occurs after brain injuries. Although there are different types of aphasia, the difficulty to recall words or names is the most common language disorder presented by aphasic individuals. In fact, it can be the only residual defect after rehabilitation of aphasia [2]. Several studies about aphasia have demonstrated the positive effect of speech therapy activities for the improvement of social communication abilities [3]. Typically, word retrieval problems can be treated through word naming therapeutic exercises. In fact, it has been shown that the frequency and the intensity of speech therapy are key factors in the recovery of lost communication functionalities in aphasic patients [4].

Recently, we have presented the first prototype of an on-line platform that incorporates speech and language technology (SLT) for treatment of Portuguese aphasic speakers with lost or reduced word naming ability [5]. The system –named VITHEA (Virtual Therapist for Aphasia treatment)– consists of a web-based platform that permits speech therapists to easily create speech therapy exercises that can be later accessed by patients using a web-browser (see Fig. 1). During the training sessions, the role of the speech therapist is taken by a "virtual therapist" that presents the exercises and that is able to validate the patients answers by means of the use of automatic speech recognition (ASR). Thus, the platform makes available word naming exercises to aphasia patients from their homes at any time, which will certainly allow an increase in the number of training hours, and hopefully a significant improvement in the rehabilitation process. In fact, in addition to serve as a complement to conventional speech therapy sessions, the system can be of great benefit for therapists as a tool to assess and track the evolution of their patients.

This work was funded by the FCT project RIPD/ADA/109646/2009, and partially supported by FCT (INESC-ID multiannual funding) through the PIDDAC Program funds.

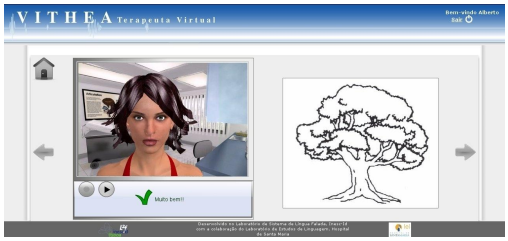

Figure 1: Screen-shot of the VITHEA patient application.

In this work, we focus on the description and assessment of our built-in automatic word naming recognition module and on its role within the VITHEA system. The method proposed to validate or reject patients' answers is based on a keyword spotting approach that makes use of a model of background speech in competition with the expected target keyword model. Given that our in house speech recognition engine is based on the hybrid HMM/MLP paradigm [6], the problem of deriving a robust background speech model is specifically addressed. Thus, we derive a simple solution based on the computation of the likelihood of the competing phonetic classes that does not need acoustic model re-training. Moreover, since targeted users are patients with word naming difficulties but without articulatory or speech production impairments, there is no need for acoustic model adaptation and general purpose acoustic models can be used. A corpus consisting of word naming exercises during ordinary speech therapy sessions of aphasia patients has been collected to evaluate the proposed word naming detector. The reliability of the word naming detector for both global evaluation and training purposes is then investigated. Finally, a calibration method to adapt the VITHEA system to the type of speech, acoustic conditions and the recovery stage of each patient is proposed.

2. Automatic word naming recognition

2.1. Definition of word verification task

The targeted task for automatic word naming recognition consists of deciding whether a claimed word W is uttered in a given speech segment S or not. We refer to this task as word verification. In the simplest case, a true/false answer is provided, but a verification score might be also generated. Notice that we name it word verification although a keyword may in fact consist of more than one word.

2.2. Word verification based on keyword spotting

Several approaches exist based on speech recognition technology to tackle the word verification problem. Given that word W is known, forced alignment with an automatic speech recognition system could be one of the most straightforward possibilities. However, we expect that speech from aphasic patients will contain a considerable amount of hesitations, doubts, repetitions, descriptions and other speech disturbing factors that can make this approach inconvenient. Alternatively, keyword spotting (KWS) methods [7] can better deal with unexpected speech effects. Concretely, KWS based on acoustic matching of speech with keyword models in contrast to a competing model –generally known as background, garbage or filler speech model– is an adequate solution for the type of problem addressed in the on-line therapy system.

2.3. Keyword spotting with AUDIMUS

The in-house ASR engine named AUDIMUS, that has been previously used for the development of several ASR applications such as the recognition of Broadcast News (BN) for several languages [8], has been integrated into the VITHEA system. In order to do so, the baseline ASR system was modified to incorporate a competing background speech model that is estimated without the need for acoustic model re-training.

2.3.1. The baseline speech recognizer

AUDIMUS is a hybrid recognizer that follows the connectionist approach [6]. The baseline system combines three MLP outputs trained with Perceptual Linear Prediction features (PLP, 13 static + first derivative), log-RelAtive SpecTrAl features (RASTA, 13 static + first derivative) and Modulation SpectroGram features (MSG, 28 static). The version of AUDIMUS integrated in VITHEA uses a general purpose gender independent acoustic model trained with 57 hours of downsampled Broadcast News data and 58 hours of mixed fixed-telephone and mobile-telephone data in European Portuguese [9]. The number of context input frames is 13 for the PLP and RASTA networks and 15 for the MSG network. Neural networks are composed by two hidden layers of 1500 units each one. Monophone units are modelled, which results in MLP networks of 39 soft-max outputs (38 phonemes + 1 silence). For the word naming detection task, an equally-likely 1-gram language model formed by the possible target keywords and a competing speech background model is used. The minimum duration for the background speech word is set to 250 msec. The AUDIMUS decoder is based on a weighted finite-state transducer (WFST) approach to large vocabulary speech recognition.

2.3.2. Background modelling with HMM/MLP recognizer

While keyword models are described by their sequence of phonetic units provided by an automatic grapheme-to-phoneme module, the problem of background speech modelling must be specifically addressed. The most common method consists of building a new phoneme classification network that in addition to the conventional phoneme set, also models the posterior probability of a background speech unit representing "general speech". This is usually done by using all the training speech as positive examples for background modelling and requires re-training the acoustic networks. Alternatively, the posterior probability of the background unit can be estimated based on the posterior probabilities of the other phonetic classes. For instance, as the mean probability of the top-N most likely outputs of the phonetic network at each time frame [10]. Here, we followed a similar approach to the latter, but instead of computing the posterior probability of the background speech as an average in the probability domain, we have computed the average in the likelihood domain. That is, the likelihood of feature vector $\mathbf{x}$ for phonetic class c_l is

$$p(\mathbf{x}|c_l) \propto \frac{p(c_l|\mathbf{x})}{p(c_l)} \tag{1}$$

where $p(c_l)$ is the class prior and $p(c_l|\mathbf{x})$ the class posterior modelled with the MLP. Then, the likelihood of the background speech model can be computed as

$$p(\mathbf{x}|c_{bg}) = \tfrac{1}{N} \sum_{l=1}^{N} p(\mathbf{x}|{c_l}') \tag{2}$$

where $\mathbf{c}' = \{{c_1}' \dots {c_l}' \dots {c_L}'\}$ is the list of sorted classes from maximum to minimum likelihood and N is the number of top most likely non-silence classes (silence is excluded from the average computation). Notice that, since the decoder operates in the likelihood domain, there is not need for estimating the prior of the background class, in contrast to the approach that computes the average in the posterior probability domain. In practice, although there are not large differences between the likelihood based method and the posterior probability based approach, we have observed a smoother KWS performance in the former case when varying the background tuning parameter described in next section. In this work, we used the top-6 most likely classes for background speech likelihood computation.

2.3.3. Background scale for word spotting tuning

In order to control the weight of the background speech competing model with respect to the keyword models in the decoding process, we have introduced a background scale term β in the computation of the acoustic score of the background phonetic class. The term is exponential in the likelihood domain, that is linear in the acoustic score (log-likelihood) domain. This background scale permits adjusting the word naming detection system to penalize the background speech model or to favour it. In this way, it is possible to make the system more prone towards keyword detections (and possibly false alarms) or towards keyword rejections (and possibly miss detections).

3. Aphasia Portuguese Speech corpus

A database recorded during regular therapy sessions of native Portuguese speakers with different types and degrees of aphasia has been collected. Each of the sessions consisted of naming exercises with 103 objects presented at intervals of at most 15 seconds. The objects and the presentation order were the same for all patients. Recordings took place in two different speech therapy centres in two phases. In both phases, the same common set-up was used for recording, which consists of a lap-top computer and two inexpensive microphones: a built-in head-set microphone and a table-top microphone. Inexpensive microphones were preferred to high-quality ones in order to better resemble the actual speech recordings of potential users. With the same intention, background noise conditions were not particularly controlled. Data originally captured at 44.1 kHz was downsampled to 8 kHz to match the acoustic models sampling frequency. Segmentation and word-level transcriptions were manually produced for each session. From the complete manual transcriptions, only the excerpts that correspond to patient's

answers to naming exercises were extracted and used for evaluation. A word naming exercise was considered correct whenever the targeted word was said by the patient (independently of its position, amount of silence before the valid answer, etc...). It is worth noticing that this is not necessarily the criteria followed in therapy tests by speech therapists, where doubts, repetitions, corrections, approximation strategies and other similar factors can be indicators of speech pathologies. The detection of these speech artefacts is not addressed in this work, although it is an extremely appealing problem for future research.

APS-I The first phase of data collection campaign was carried out in February and March of 2011. It includes speech from 8 aphasia patients recorded in a small office room of wooden walls. The total number of extracted segments for evaluation is 1004 with approximately 1 hour and 30 minutes of total duration.

APS-II The second data collection was carried out during May and June of 2011. The original set consisted of 18 new patients recorded in a larger office room. Unfortunately, during the manual transcription phase it was found that the audio quality was severely affected by the presence of electrical noise. In order to partially reduce this effect, a noise removal processing was applied to the recorded data. Nevertheless, the resulting speech quality was still poor (much worse than APS-I) and only the recordings of the patients with a subjective good speech quality were kept. Finally, the APS-II data set includes speech from 8 aphasia patients with 850 extracted segments for evaluation and a total duration of 63 minutes.

APS-I				APS-II			
Patient	Gender	Age	WNS	Patient	Gender	Age	WNS
1	m	57	0.60	9	m	44	0.86
2	f	74	0.55	10	f	34	0.36
3	m	65	0.33	11	m	48	0.69
4	m	60	0.22	12	m	21	0.16
5	m	78	0.75	13	m	62	0.84
6	f	52	0.84	14	f	24	0.75
7	f	57	0.63	15	m	77	0.64
8	m	52	0.93	16	f	19	0.98

Table 1: APS-I and APS-II patients information including gender, age and word naming score (WNS) computed in the APS data-set.

4. Experimental evaluation

The APS corpus is used to evaluate the automatic word naming recognition system. Two metrics are considered throughout this section: the word naming score (WNS) and the word verification rate (WVR). On the one hand, the WNS is computed for every speaker as the number of positive word detections divided by the total number of exercises. The manual or reference WNS is shown in Table 1. The automatic WNS in contrast to the manual WNS can be considered a measure of the goodness of the word detector as a tool for global evaluation of patients' word naming ability. On the other hand, the WVR is computed for every speaker as the number of coincidences between the manual and automatic result (true acceptances and true rejections) divided by the total number of exercises. Thus, it is a measure of the reliability of the detector as a verification tool for virtual word naming therapy exercises. In these experiments, only the head-set microphone recordings have been considered.

4.1. Baseline and oracle word naming results

The baseline configuration of the automatic detector is used in this first set of experiments. The leftmost plot of Figure 2 shows the WNS for each patient of APS-I (1 to 8) computed by a human evaluator (blue bar) and by the automatic detector (red bar). In this case, the Pearson's correlation coefficient between human and automatic scores is 0.904. Moreover, the average absolute difference between the human and automatic scores is 0.074. This means that both scoring methods provide similar figures for each patient, besides providing highly correlated scores. The leftmost side of Table 2 shows the WVR for every speaker of APS-I and the average WVR. These results can be generally considered quite promising. The automatic word detector for this data set is able to provide significant global evaluation results, but also high word verification rates which would permit performing reliable word naming therapy exercises. On the other hand, the rightmost plot of Figure 2 shows manual and automatic WNS for each patient of APS-II (9 to 16). In contrast to the results obtained with the APS-I patients, there is a strong degradation of the automatic scores with this data set that results in an average absolute difference between human and automatic scores of 0.314. Furthermore, the WVR shown in the rightmost side of Table 2 for every speaker is quite low and the average WVR is reduced to 0.663 (in contrast to APS-I 0.804).

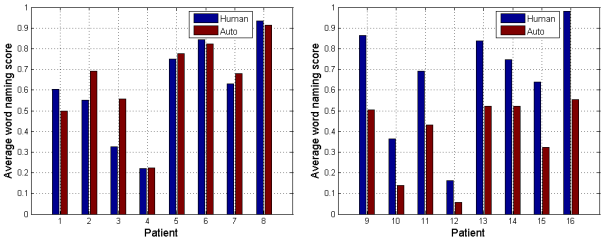

Figure 2: On the left side, average word naming scores of the human and automatic evaluations for the APS-I corpus. On the right side, average word naming scores of the human and automatic evaluations for the APS-II corpus.

APS-I		APS-II	
Patient	WVR	Patient	WVR
1	0.78	9	0.64
2	0.80	10	0.66
3	0.73	11	0.64
4	0.84	12	0.87
5	0.73	13	0.63
6	0.91	14	0.70
7	0.70	15	0.58
8	0.91	16	0.57
avg	0.80	avg	0.66

Table 2: Word verification rate (WVR) for the patients of APS-I and APS-II data sets and average WVR, using the baseline automatic word verification system.

The background scale term (β) described in section 2.3.3 can be used as a method to adjust the word detector to the speech production characteristics of each patient and to the particularities of the recorded audio. Table 3 shows WVR results in the

case in which the optimum background scale term for each patient (the one that gives a higher WVR) is known. These results confirm that it is possible to find an operation point of the automatic word detector adapted to the type of speech and to the acoustic characteristics of the data in order to obtain high word verification performances. The remaining problem is how to select this background scale for a certain set of word naming exercises of a particular patient. An hypothesized solution would be adjusting the word detector according to a parameter that characterizes the word naming ability of each patient, for instance, the manual WNS. We verified that if the β term used for each patient is selected in order to minimize the absolute difference between the manual and the automatic WNS, the obtained average WVR increases to 0.824 and 0.803 for the APS-I and APS-II data sets respectively.

APS-I		APS-II	
Patient	WVR	Patient	WVR
1	0.84	9	0.90
2	0.84	10	0.74
3	0.83	11	0.76
4	0.86	12	0.91
5	0.75	13	0.83
6	0.94	14	0.77
7	0.72	15	0.73
8	0.94	16	0.91
avg	0.84	avg	0.82

Table 3: Word verification rate (WVR) for the patients of APS-I and APS-II data sets and average WVR, using ideal background scale terms for each patient.

4.2. Calibration of automatic word naming verification

The previous results suggest that an automatic computation of the β parameter that provides an automatic WNS closer to the manual value might be used as a calibration criteria. In order to validate this hypothesis, the data from every speaker was randomly split into two halves, in a cross-validation experiment. The first half is used to search for the best β parameter on that data sub-set. Then, this β scale term is used to process the second half of the data and the WVR is computed on this second sub-set. Note, however, that if we consider a single partition, a bias may be introduced. Hence, we repeat this process 10 times with different random partitions of the two halves of data. Mean word verification rates and standard deviation for each patient are shown in Table 4.

Regarding the baseline results of Table 2, a slightly better average performance is obtained in the APS-I corpus using the proposed calibration method, but more importantly, a large improvement is achieved with the APS-II patients. This is an important result for the VITHEA system. If the patients word naming ability can be characterized by speech therapists based on similar exercises to the ones proposed by the virtual therapist, the system can use this characterization to adapt the word detector to any particular patient after a certain number of word naming exercises. Moreover, the calibration method permits adapting to different acoustic conditions, such as the ones presented by APS-I and APS-II data. In fact, this method can be used to adapt regularly to the word naming ability progress of each patient measured by speech therapists, which is expected to evolve as a result of the speech therapy sessions.

APS-I		APS-II	
Patient	WVR	Patient	WVR
1	0.84 ($\pm$0.036)	9	0.86 ($\pm$0.032)
2	0.84 ($\pm$0.020)	10	0.74 ($\pm$0.046)
3	0.83 ($\pm$0.038)	11	0.70 ($\pm$0.046)
4	0.82 ($\pm$0.024)	12	0.89 ($\pm$0.039)
5	0.70 ($\pm$0.024)	13	0.78 ($\pm$0.044)
6	0.93 ($\pm$0.030)	14	0.74 ($\pm$0.059)
7	0.69 ($\pm$0.063)	15	0.73 ($\pm$0.035)
8	0.92 ($\pm$0.032)	16	0.92 ($\pm$0.036)
avg	0.82	avg	0.795

Table 4: Word verification rate (WVR) for the patients of APS-I and APS-II data sets and average WVR, using automatically calibrated background scale terms.

5. Conclusions

In this work we have investigated the use of automatic word naming recognition as part of an on-line treatment system aimed at acting as a virtual therapist for word naming ability training. Using a new collected database of word naming exercises performed by native Portuguese speakers with aphasia, we have shown that it is possible to achieve highly correlated global word naming scores and high performance word verification rates even for different types of patients and acoustic conditions. In general, we consider the achieved results very promising and in the near future we plan to evaluate the utility of the whole therapy platform in terms of word naming recovery progress.

6. References

[1] Pedersen, P.M., Jorgensen, H.S., Nakayama, H., Raaschou, H.O. and Olsen, T.S., "Aphasia in acute stroke: incidence, determinants, and recovery", Ann. Neurol. 38, pp. 659-666, 1995.

[2] Wilshire, C.E. and Coslett, H.B., "Disorders of word retrieval in aphasia theories and potential applications", In S. Nadeau, L.J.G. Rothi e B. Cronon (Eds.), Aphasia and Language. Theory to practice, pp. 82-107, New York: The Guilford Press, 2000.

[3] Basso, A., "Prognostic factors in aphasia", Aphasiology, 6 (4), pp. 337-348, 1992.

[4] Bhogal, S.K., Teasell, R. and Speechley, M., "Intensity of aphasia therapy, impact on recovery", Stroke, pp. 987-993, 2003.

[5] Pompili, A., Abad, A., Trancoso, I., Fonseca, J., Martins, I.P., Leal, G., Farrajota, L., "An on-line system for remote treatment of aphasia", Proc. Second Workshop on Speech and Language Processing for Assistive Technologies, 2011.

[6] Morgan, N. and Bourlard, H., "Continuous Speech Recognition: An introduction to the Hybrid HMM/connectionist approach", IEEE Signal Processing Magazine, 12(3):25–42, 1995.

[7] Szöke, I., Schwarz, P., Matějka, P., Burget, L., Karafiát, M., Fapšo, M. and Černocký, J., "Comparison of Keyword Spotting Approaches for Informal Continuous Speech", Proceedings of Interspeech'05, pp 633-636, 2005.

[8] Meinedo, H., Abad, A., Pellegrini, T., Trancoso, I. and Neto, J. "The L2F Broadcast News Speech Recognition System", In Fala2010, Vigo, Spain, 2010.

[9] Abad, A. and Neto, J., "Automatic classification and transcription of telephone speech in radio broadcast data", In International Conference on Computational Processing of Portuguese Language (PROPOR 2008), Portugal, 2008.

[10] Pinto, J., Lovitt, A. and Hermansky, H., "Exploiting Phoneme Similarities in Hybrid HMM-ANN Keyword Spotting", Proceedings of Interspeech'07, Aug 2007.