AUTOMATIC RECOGNITION OF PROSODIC PATTERNS IN SEMANTIC VERBAL FLUENCY TESTS - AN ANIMAL NAMING TASK FOR EDUTAINMENT APPLICATIONS

H. Moniz[1,2], A. Pompili[1,4], F. Batista[1,3], I. Trancoso[1,4], A. Abad[1,4], C. Amorim[1,4]

[1]Spoken Language Systems Lab - INESC-ID, Lisbon, Portugal

[2]FLUL/CLUL, Universidade de Lisboa, Portugal

[3]ISCTE - Instituto Universitário de Lisboa, Portugal

[4]IST, Lisboa, Portugal

helena.moniz,anna.pompili,fernando.batista,isabel.trancoso,alberto.abad, cristiana.amorim@l2f.inesc-id.pt

ABSTRACT

This paper automatically detects prosodic patterns in the domain of semantic fluency tests. Verbal fluency tests aim at evaluating the spontaneous production of words under constrained conditions. Mostly used for assessing cognitive impairment, they can be used in a plethora of domains, as edutainment applications or games with educational purposes. This work discriminates between list effects, disfluencies, and other linguistic events in an animal naming task. Recordings from 42 Portuguese speakers were automatically recognized and AuToBI was applied in order to detect prosodic patterns, using both European Portuguese and English models. Both models allowed to differentiate list effects from the other events, mostly represented by the tunes: L* H/L(-%) (English models) or L*+H H/L(-%) (Portuguese models). However, English models proved to be more suitable because they rely in substantial more training material.

Keywords: Prosody, Semantic Fluency, Edutainment, and Automatic Speech Recognition

1. INTRODUCTION

Tests based on the verbal generation of words are widely used in several domains, from language disorder or neuropsychological assessment [5, 6, 11] to edutainment applications [23, 7, 24, 12]. There are, typically, two classes of these tests: one in which subjects are invited to generate a single response word to a provided target, and one in which subjects are invited to produce exemplars from a specified category. The former is referred as *generation* or *naming* task, and the latter as *fluency task*. Fluency tasks are useful because they assess both associative processes and strategic abilities. The semantic verbal fluency test belongs to this category, aiming at evaluating the spontaneous production of words within specified semantic categories and under constrained conditions.

Semantic fluency tests are specially important for cognitive impairment diagnosis and monitoring, but they may be also interesting in the context of edutainment. Research has shown that cognitive skills, which can fade without stimulation as we age, can be improved by playing brain games [4]. Besides cognitive abilities, exercises and brain games can benefit both verbal and short-time memory skills. In this context, there exist, already, a plethora of games that, similarly to semantic fluency tasks, aim at stimulating verbal fluency and mental processing speed or yet at encouraging word finding and helping to enrich a person's vocabulary. In these tests, the user is asked to name as many words as possible belonging to a specific category and within a time constrained interval, typically one minute. The most common category is the *animals'* category [34], though there are other common categories, such as food or first names. In animal naming tests, the score corresponds to the sum of all admissible words, where names of extinct, imaginary or magic animals are considered admissible, while inflected forms and repetitions are considered inadmissible.

Automatic Speech Recognition (ASR) could be of valuable support in the automation of tests in naming and fluency tasks. However, while the implementation of the first tests is nowadays feasible with current ASR technology, the automation of semantic fluency tests still raises several challenges. One of the challenges in the automation of such tests is the vastness of the domain addressed by the task. This implies the availability of a complete source of knowledge over the domain of the semantic category chosen and a sensible method for restricting the vocabulary size. In fact, in order to improve recognition accuracy, the vocabulary should contain only those items classified as the most popular terms. The size of this list may have a significant impact on the outcome of the recognizer. If a keyword is missing from the list, it will never be detected, on the other hand we also expect that a longer list will result in an increase of the model perplexity.

Another challenge for automatic word naming exercises is the presence of disfluencies, which may significantly affect the performance of current speech recognition systems. Disfluencies are relatively frequent in spontaneous speech, but in this context they may be particularly relevant, because of the cognitive load required by the test and it's duration. The false alarms that may be triggered by disfluencies and comments by the speaker are the major goal of this study. In particular, we want to investigate the prosodic patterns of lists *vs.* other events,

given that animal names in this type of exercises are typically pronounced with list effects.

This paper describes our work in improving the detection of words belonging to lists based exclusively on their prosodic patterns. Our first approach to this task was therefore aimed at establishing the distinctive prosodic patterns used by healthy adults in animal naming lists, targeting an edutainment scenario as a first step, that will later be extended to eHealth applications. The use of prosodic features can help predict more accurately correctly recognized turns in dialogue systems [13], rather than the use of acoustic confidence scores alone. It is also known that list effects or serial recall tasks [30, 29] display prosodic characteristics mostly characterized by two patterns: i) a continuation rise contour, a rising F0 movement from the nuclear or prenuclear syllables up to the end of the phrase; and ii) a finality contour, a fall from the nuclear or prenuclear syllables until the end of the phrase. The continuation contour express that the list is to be continued and the finality contour that the item is the last one of a recall series or the last one in the entire file. Building on that, we hypothesized that by applying AuToBI [26, 27] models we would be able to capture the prosodic patterns of lists and then use them to better predict words produced in an animal naming task. Having this hypothesis confirmed, we can further encompass other semantic categories based on generic language models combined with automatic prosodic patterns detection.

This paper is organized as follows. Section 2 overviews the collected corpus. Experiments and most relevant results regarding ASR and AuToBI will be described in Section 3. Finally, Section 4 describes core conclusions and future trends.

2. CORPUS

The corpus used in this paper is a database of recordings of native Portuguese speakers that has been collected for the assessment of semantic verbal fluency tests, at this point an animal naming task. It is composed of 42 healthy adults (19 females and 23 males), with ages varying from 20 to 65 years, different education, socioeconomic status, and cultural background. The corpus was collected with the aim of having a diverse sample of adults in an animal naming task in a first stage, aiming at encompassing distinct categories as in an encyclopedia game (countries, fruits, surnames, *inter alia*). The recordings took place in several sessions, with different conditions. Three different microphones have been used, two head-set and a laptop microphone. No particular constraint over background noise condition was imposed. Each of the sessions consisted approximately of one-minute recordings, in which the subject was asked to name the animals he/she was able to remember within the available time. Data originally captured at 16 kHz was down-sampled to 8 kHz to match the acoustic models sampling frequency. Orthographic transcriptions were manually produced for each session, and all the events were classified as a word from an animal list, as a disfluency or as other events, namely comments.

The overall duration of the corpus is approximately 43 minutes, of which about 21 minutes are silent, about 15 minutes include speech, and the remaining ones contain disfluencies and other paralinguistic events (*e.g.*, laugh, cough) or background noise. Disfluencies include mostly filled paused (*aa*/uh; *aam, eem*/um), fragments (*ca- cavalo*/hor- horse), and prolongations (*de*=/of pronounced as [d@:]). Occurrences of repetitions of an animal name already produced were only taken into account when the speaker is aware of this and uses explicit editing terms, *e.g.*, *"já disse"/"*I already said it*"*. The total number of words uttered by all the subjects, including Out-of-Vocabulary (OOV) words and excluding disfluencies, is 1225. The definition of OOV for the animal naming task extends to each word that does not represent an animal name, encompassing distinct types, *i.e.*: interjections "*Ah, já sei!*"/"Ah, I know"; expressions of uncertainty *"acho eu"*/"I think"; words such as or "espetada"/"skewer" or even made-up words such as "perdiniz" for "perdiz"/partridge. Thus, the number of valid words, excluding OOV, is 1171, while the number of disfluencies is 321, representing 27% of the whole corpus. This percentage is clearly not in line with the ones reported by [14, 31, 9, 8, 32, 18, 19], who indicate an interval of 5% to 10% of disfluencies in human-human conversations. This very high disfluency rate is interpretable by task effects, in particular naming animals under strict temporal constraints.

Table 1: Speech corpus properties.

	Train set	Test set	Total
Speech time	≈11 min.	4 min.	15 min.
Silent pauses	≈15 min.	≈6 min.	21 min.
Valid words	864	307	1171
Disfluencies	255	66	321

The data has been divided into two subsets: the recordings of 31 speakers, corresponding to 75% of the corpus, have been used for training the animal naming task, while the remaining ones have been used for testing the achieved results. Table 1 reports detailed information about the speech corpus, namely: total time of speech and silence, number of valid words, and number of disfluencies for the training and test corpus.

3. EXPERIMENTS AND RESULTS

The ASR experiments here described use Audimus [17], a large vocabulary continuous speech recognition module. The ASR module is integrated in the fully automatic subtitling system that is running on the main news shows of the public TV channel in Portugal, since 2008 [22]. Although the system can be improved by using dynamic vocabulary and domain-specific language models, all the experiments reported in this paper use a generic language model for broadcast news, encompassing 100k words. A possible way of tackling the animal naming task would be to use an ASR system with a generic language model. We have performed such a baseline experiment using our ASR system and, as expected, due to challenges described in the first section, the performance is very low, as shown in the first line of Table 2.

Table 2: WER for different language models.

Language Model	Train set	Test Set	Total
Generic ASR system	88.95	105.47	130.42
Initial	16.80	21.22	17.97
Ontology based	11.94	19.94	13.64

The remainder of this section presents our efforts towards the automatic detection of prosodic patterns [26, 27, 28, 33] in words produced in the context of the animal naming task. We have tested two main approaches: the first approach concerns keyword spotting and relies on two different lists of animal names that were used in order to automatically recognize animal occurrences. The second approach uses AuToBI to identify potential animal segments, based on different automatic segmentations.

3.1. Keyword spotting

Our first approach exploited a technique known as keyword spotting, already proved to be appropriate for dealing with naming tasks and also for filtering speech disfluencies [21, 3, 25]. Keyword spotting aim at detecting a certain set of words of interest in the continuous audio stream. This is achieved through the acoustic match of speech with keyword models in contrast to a background model (everything that is not a keyword). In this approach, the keyword model contains the names of admissible animals that will be accepted by the speech recognition system. The size of this list may have a significant impact on the outcome of the recognizer. In fact, if a keyword is missing from the list, it will never be detected; on the other hand longer lists will increase the perplexity of the keyword model.

In order to evaluate the success of this approach we have used two distinct metrics: i) count the number of animal names returned by the keyword model, even if the names do not match; and ii) the well-known WER (Word Error Rate).

3.1.1. Initial language model

Our baseline model consists of an existing list of animal names [15] that includes 6044 animal names, grouped, classified, and labeled with its semantic category, without inflected forms.

Considering that some names are more likely than others, we have computed the likelihood of the different target terms, as it is commonly done in n-gram based language modeling. For this purpose, the total number of results provided by any web search engine for a particular term can be used. However, incorrect counts can be obtained because of existing homonyms, which correspond to alternative meanings of the term. For that reason, we have used a refined retrieval strategy that takes into account the semantic information associated to each key term, consisting of the search query composed by the bigram: <*animal name*> <*category*>. The Bing Search API was used for obtaining the counts used in this work. The obtained values made it possible to sort the list, and we observed that the most exotic names were moved to the end. It was then possible to reduce the list to the most probable names, by setting thresholds. It was observed that shorter lists led to an expected increase in the number of misses and substitutions for some users, but it was beneficial for some other users, whose speech recognition results had shown a high number of insertions with the original list. A detailed analysis of the recognition results with shorter lists showed that many of the still existing insertions were due to the existence of a considerable amount of unusual animal names (mostly short names such as "anu" and "udu"). After evaluating several lexical and semantic resources, we used Onto.PT [10], an ontology for the Portuguese language, to filter the elements of the list also based on its content. By excluding all words that were missing in Onto.PT, the word list was reduced to the most popular 802 names and also reduced the average error up to 2%.

The second line of Table 2 reports the results for this language model in terms of WER. The keyword spotting approach returned 324 animals in the test set, plus 8 segments marked as *background* and 25 other segments marked as *filled pauses*. The keywords model returned 297 words that matched the words in the reference. The ASR system returned only 150 words that math the list used in the keyword spotting, mostly due to out-of-vocabulary words and also due to the language models not being suitable for processing lists of words.

3.1.2. Ontology based language model

Because our initial language model was too broad, we focused our research towards better resources that could provide us with structured data over different semantic domains. We have used TemaNet [16], a lexical-conceptual networks (wordnets) for the Portuguese language organized in semantic domains. In the current version, TemaNet includes twelve domains. Lexical concepts are linked by relationships of various kinds: synonymy, hyponymy/hypernymy and meronymy/holonymy. For our purposes, we exploited the hyponymy relation extracting the subtypes of animal, a subtype of the domain "Living Things". Temanet is of particular interest for our task because it is highly structured. In fact, the hyponymes of animal are organized in a hierarchy of several layers that include, among others, the separation between male and female. This is relevant not only because in Portuguese, unlike English, there are different words to express the genre of an animal (i.e.: "cão", "cadela" / dog), but also because the animal naming task evaluation rules require to account for genre difference. Our first approach with this resource, however, did not impose constraints on the deep of the hierarchy or on the type of the information extracted. We accepted all the subtypes of animal, leading to a language model composed of 400 keywords. As for the baseline language model, we have then computed the likelihood of the target terms exploiting the total number of results provided by a web search engine. Unfortunately, experiments reported a reduction of the average WER up to 4% for the train corpus, and only of 2% for the test corpus. The third line of Table 2 summarizes the results achieved.

Table 3: Performance using Temanet and ASR segmentations with AuToBI.

Segmentation	AuToBi model	Acc.	Correct
Temanet-based	PT	72.7%	314/432
	EN	84.3%	364/432
ASR generic	PT	71.8%	234/307
	EN	89.1%	298/307
Phone-based	PT	77.8%	267/343
	EN	**91.8%**	315/343

3.2. AuToBI with different automatic segmentations

The automatic generation of the keyword model requires some laborious pre-processing steps and implies having word lists for each semantic category. These facts limit the extensions of the animal naming task to other semantic categories. In order to overcome this limitation, we have also followed an alternate approach where an automatic segmentation is produced and then used together with AuToBI to identify potential animal names. This sections reports on experiments with different automatic segmentations, namely: Temanet-based keyword spotting, ASR-based, and phone-based. Phone-based segmentation is provided by a phone recognizer, which is based on the Audimus Multilayer Perceptron (MLP) outputs[1].

The different segmentations have been used as input to AuToBI, the **Au**tomatic **ToBI** annotation system (AuToBI) for Standard American English (SAE) by [26, 27]. AuToBI is a publicly available tool, which detects and classifies prosodic events following SAE intonational patterns. AuToBI relies on the fundamentals of the ToBI system, meaning it predicts and classifies tones and break indices using the acoustic correlates - pitch, intensity, spectral balance and pause/duration. Previous work on prosodic event detection using AuToBI [26, 27, 28, 33, 20] have shown that prominence and phrase boundaries can be predicted in a cross-language context (American English, German, Mandarin Chinese, Italian and French), albeit with language specific properties. Those studies also found little support for the hypothesis that language families are useful for cross-language prosodic event identification.

First, we have used AuToBI with English models in order to predict prosodic patterns (detection and classification of pitch accents and boundary tones). These models (v1.3) include training material from three corpora of read and spontaneous speech: Boston Directions Corpus, Boston University Radio News Corpus, Columbia Games Corpus ([26] and references therein). Second, we have used previously trained models for European Portuguese, based on spontaneous and prepared speech corresponding to a small dataset of about 33 minutes [20].

Table 3 summarizes the results achieved with the different segmentations. By applying the English models with the generic ASR language model, the system correctly identifies 298 speech segments as potential animal names, with an accuracy of 89.1%. The achieved result is very similar to the 297 that is the number of matching words between the reference and the keywords model, which suggests that this approach is suitable for identifying potential animal names. Moreover, because it is task-independent, this approach can be easily ported to other semantic categories, providing motivation to use the system for edutainment purposes, for an encyclopedia game, for instance. With the Portuguese prosodic models, the system correctly classifies 234 speech segments as potential animal names, with 71.8% accuracy. Such decrease in performance may be interpreted by the fact that the Portuguese models used significantly less training data than the English ones. The best performance is achieved with the phone recognizer also using AuToBI English models with an accuracy of 91.8%.

Both English and Portuguese models allowed for the differentiation of list effects from the other events, mostly represented by the tunes: L* H/L(-%) (English models) or L*+H H/L(-%) (Portuguese models). Two main challenges still endure: sequences of two or more animals produced as a serial group with coarticulation, in which the last animal is the only item marked with the continuation meaning contour; and terminality contours associated either with an animal or with a comment by the speakers, as "já disse" (already said it).

4. CONCLUSIONS

This paper reports our preliminary work on detecting prosodic patterns in semantic fluency tests, concretely in an animal naming task for edutainment purposes, such as the creation of an encyclopedia game. This work corresponds to one of the first uses of AuToBI models for Portuguese. We applied AuToBI models to capture the prosodic patterns of lists and then use them to predict potential animal names. AuToBI models for European Portuguese and English were applied, in order to verify if prosodic patterns could be identified. Both models allowed for the differentiation of list effects from the other events, mostly represented by the tunes: L* H/L(-%) (English models) or L*+H H/L(-%) (Portuguese models). The English models proved to be more suitable for the task since they rely on substantial training material. Results are quite encouraging and may be considered as a step-forward towards the use of other semantic categories, in order to account for edutainment applications such as an encyclopedia game.

We plan to integrate the developed method into an online platform that will allow for the automatic fruition of verbal fluency tasks in different domains. First, semantic fluency tasks will be implemented for providing an automatic screening assessment for cognition disorder, then they will be adapted for extending an existing therapy system devoted to aphasia rehabilitation [25, 2, 3].

ACKNOWLEDGMENTS

This work was supported by national funds through – Fundação para a Ciência e a Tecnologia, under Grants SFRH/BPD/95849/2013 and SFRH/BD/97187/2013, funds with reference UID/CEC/50021/2013, and by EU-IST FP7 project SpeDial under contract 611396.

5. REFERENCES

[1] Abad, A., Neto, J. 2008. Incorporating acoustical modelling of phone transitions in an hybrid ANN/HMM speech recognizer. *Interspeech*.

[2] Abad, A., Pompili, A., Costa, A., Trancoso, I. 2012. Automatic word naming recognition for treatment and assessment of Aphasia. *13th Annual Conference of the International Speech Communication Association (InterSpeech 2012)*.

[3] Abad, A., Pompili, A., Costa, A., Trancoso, I., Fonseca, J., Leal, G., Farrajota, L., Martins, I. P. 2013. Automatic word naming recognition for an on-line Aphasia treatment system. *Computer Speech & Language* 27(6), 1235–1248. Special Issue on Speech and Language Processing for Assistive Technology.

[4] Anguera, J. A., Boccanfuso, J., Rintoul, J. L., Al-Hashimi, O., Faraji, F., Janowich, J., Kong, E., Larraburo, Y., Rolle, C., Johnston, E., Gazzaley, A. 2013. Video game training enhances cognitive control in older adults. *Nature* 501, 97–101.

[5] Arroyo-Anlló, E., Lorber, M., Rigaleau, F., Gil, R. 2012. Verbal fluency in Alzheimer's disease and Aphasia. *Dementia* 11(1), 5–18.

[6] Binetti, G., Magni, E., Cappa, S. F., Padovani, A., Bianchetti, A., Trabucchi, M. 1995. Semantic memory in Alzheimer's disease: An analysis of category fluency. *Journal of Clinical and Experimental Neuropsychology* 17(1), 82–89.

[7] Charsky, D. 2010. From edutainment to serious games: a change in the use of game characteristics. *Games and Culture* 5, 177–198.

[8] Clark, H. H. 1996. *Using language*. Cambridge University Press.

[9] Fox-Tree, J. E. 1995. The effects of false starts and repetitions on the processing of subsequent words in spontaneous speech. *Journal of Memory and Language* (34), 709–728.

[10] Gonçalo Oliveira, H., Gomes, P. August 2010. Onto.PT: Automatic Construction of a Lexical Ontology for Portuguese. *Proceedings of 5th European Starting AI Researcher Symposium* STAIRS 2010. IOS Press 199–211.

[11] Grossman, M. 1978. The game of the name: An examination of linguistic reference after brain damage. *Brain and Language* 6(1), 112 – 119.

[12] Gupta, P. D. F. I., S. 2014. Designing serious games for cognitive assessment of the elderly. *Proceedings of the International Symposium of Human Factors and Ergonomics in Healthcare* 28–35.

[13] Hirschberg, H., Litman, D., Swerts, M. 2004. Prosodic and other cues to speech recognition failures. *Speech Communication* (43), 155–175.

[14] Levelt, W. 1989. *Speaking*. Cambridge, Massachusetts: MIT Press.

[15] Mamede, N. J., Baptista, J., Diniz, C., Cabarrão, V. 2012. String: An hybrid statistical and rule-based natural language processing chain for portuguese. *International Conference on Computational Processing of Portuguese Propor*.

[16] Marrafa, P., Amaro, R., Mendes, S., Lourosa, S., Chaves, R. P. 2006. Temanet - wordnets temáticas do português. http://www.instituto-camoes.pt/temanet.

[17] Meinedo, H., Viveiros, M., Neto, J. 2008. Evaluation of a live broadcast news subtitling system for Portuguese. *Interspeech*.

[18] Moniz, H., Batista, B., Mata, A. I., Trancoso, I. accepted. Towards automatic language processing and intonational labeling in European Portuguese. In: Henriksen, N., Armstrong, M., Vanrell, M., (eds), *Interdisciplinary approaches to intonational grammar in Ibero-Romance*. John Benjamins.

[19] Moniz, H., Batista, F., Mata, A. I., Trancoso, I. 2014. Speaking style effects in the production of disfluencies. *Speech Communication* 65(4), 20–35.

[20] Moniz, H., Mata, A. I., Hirschberg, J., Batista, F., Rosenberg, A., Trancoso, I. 2014. Extending AuToBI to prominence detection in European Portuguese. *Speech Prosody 2014* Ireland.

[21] Moniz, H., Mata, A. I., Viana, M. C. September 2007. On filled pauses and prolongations in European Portuguese. *Interspeech 2007* Belgium.

[22] Neto, J., Meinedo, H., Viveiros, M., Cassaca, R., Martins, C., Caseiro, D. 31 2008-April 4 2008. Broadcast news subtitling system in Portuguese. *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on* 1561–1564.

[23] Okan, Z. 2003. Edutainment: is learning at risk? *British Journal of Educational Technology* 34(3), 255–264.

[24] Pellegrini, T., Correia, R., Trancoso, I., Baptista, J., Mamede, N., Eskenazi, M. 2013. ASR-based exercises for listening comprehension practice in European Portuguese. *Computer Speech and Language* 27(5), 1127–1142.

[25] Pompili, A., Abad, A., Trancoso, I., Fonseca, J., Martins, I. P., Leal, G., Farrajota, L. 2011. An on-line system for remote treatment of Aphasia. *Proceedings of the Second Workshop on Speech and Language Processing for Assistive Technologies* SLPAT '11. Association for Computational Linguistics 1–10.

[26] Rosenberg, A. 2009. *Automatic Detection and Classification of Prosodic Events*. PhD thesis University of Columbia.

[27] Rosenberg, A. 2010. AuToBI – A Tool for Automatic ToBI annotation. *Interspeech 2010*.

[28] Rosenberg, A., Cooper, E., Levitan, R., Hirschberg, J. 2012. Cross-language prominence detection. *Proc. of Speech Prosody* Shangai, China.

[29] Savino, B. A. G. M., M. 2014. Intonational cues to item position in lists: evidence from a serial recall task. *Speech Prosody 2014* 708–712.

[30] Savino, M. 2004. *Intonational cues to discourse structure in a variety of Italian*. Tuebingen:Niemeyer.

[31] Shriberg, E. 1994. *Preliminaries to a Theory of Speech Disfluencies*. PhD thesis University of California.

[32] Shriberg, E. 2001. To "errrr" is human: Ecology and acoustics of speech disfluencies. *Journal of the International Phonetic Association* 31, 153–169.

[33] Soto, V., Cooper, E., Rosenberg, A., Hirschberg, J. 2013. Cross-language phrase boundary detection. *Proc. of ICASSP* Vancouver, Canada.

[34] Strauss, E., Sherman, E., Spreen, O. 2006. *A compendium of neuropsychological tests: Administration, norms, and commentary*. Oxford University Press.