Error diagnosis and classification of errors in two Hebrew state-of-the-art automatic speech recognition systems

Vered Silber-Varod[1], *Nitza Geri*[2]

[1] The Research Center for Innovation in Learning Technologies, The Open University of Israel
[2] Department of Management and Economics, The Open University of Israel

vereds@openu.ac.il, nitzage@openu.ac.il

Abstract

In this research we diagnose two commercial automatic speech recognizers (ASRs) on a corpus of academic lectures in Hebrew. Our goal is not only to measure the engines' performance but to find out if current Hebrew ASRs' transcription can be a reasonable replacement to human transcription, or at least a significant bootstrapping for a manual post-processing of the automatic output. We performed a word error rate (WER) diagnosis and a linguistic error classification on two automatic transcriptions – Nuance's and Google's, and compared it to a real-time (RT) stenographer's records, as well as to an exact transcription that reflects excatly the speaker's speech. Results show that the ASRs' WER is caused by massive substitutions, while the RT transcription's errors were caused mainly due to deletions. This research provides an opportunity to explore cost/benefit aspects of automatic vs. manual audio transcriptions.

Index Terms: automatic speech recognition (ASR), audio transcription, academic lectures, word error rate (WER), Hebrew

1. Introduction

Live speech is a linear mechanism of communication. Recorded speech is not linear since it enables back and forth mechanism. The challenge of audio transcription is that presently, access and flexible navigation in speech data can be achieved only by speech-to-text transformation. The text may be produced manually, by professionals, or automatically [1]. However, manual transcriptions are a difficult, time-consuming, and costly task in all languages. In under-resourced languages, such as Hebrew (with approximately 9 million speakers), this technology was therefore not fully developed for daily use. In this research we diagnose two commercial automatic speech recognizers on a corpus of academic lectures in Hebrew. Our goal is not only to measure the engines' performance but to find out if current Hebrew automatic transcription can be a reasonable replacement to human transcription, or at least a significant bootstrapping for a manual post-processing of the automatic output.

On December 5th, 2011, Google announced the launch of Voice Search in Arabic and Hebrew for Android and iPhone users. This event put the Hebrew language as one of 28 other languages that have this application [2]. In order to train Google's system, they collected over one million utterances in Hebrew, as it is used in the spoken language. The language model was up to 5 grams, and Google conducted over 250K search queries during the training phase [3]. The reported results were said to be 24.6% WER with no diacritics (Nikud). Almost a year later, on October 17th, 2012, Nuance announced that Nuance Dragon Dictation and Dragon Search Apps were available in Israel, thus expanding the apps' availability to 37 languages [4].

Transcribing natural speech, such as academic lectures is a challenge, which Automatic Speech Recognition (ASR) research had been trying to deal with for decades [5]. An overview of WER in various ASR tests compared to human perception was described in [1], where it is demonstrated how WER increases the more complex the speech material is. From less than 1% error rate, of human's and machine's, for a digit corpus, to 1.65% human error rate on continuously spoken letter and 5% machine error rate for isolated spoken letters; larger vocabulary of 1,000 words increases WER, both for human perception and machine transcription (2% and 17%, respectively). WER is also largely affected by non-linguistic conditions, such as noisy environment, variety of recording environments, sound effects and multiple speakers. Large vocabulary continuous speech recognition (LVCSR) achieved the worst machine WER results – 43%, compared to 4% of human WER. [6] has built an ASR based on TED Talks leading to a WER score of 17.4%. For Google's ASR, WER was reported as 24.8% for cross-dialect Arabic voice search [7]. In the Arabic dialect research, users spoke their search queries, typically using a mobile phone, and the system returned a transcription and web search results.

Error analysis is nowadays used for localized error detection [8], for the purpose of automatic targeting a specific mis-recognized word in an utterance in Spoken Dialogue Systems. An analysis on errors that occur in spoken SMS messages was carried by [9] who proposed an approach to detect these critical errors (i.e., errors that change the meaning of the message). They found that some errors are more important than the others, and observed that major errors often occur in a sequence, and they also occur in content words such as verbs, nouns and proper-nouns.

2. Research outline

2.1. Data

The corpus under investigation is of academic lectures genre, all were carried at the same day and during the same symposium in July, 2013. The lectures are technology-oriented and discussed different aspects of accessibility on the internet. The lecturers were 5 women and 2 men. It should be mentioned that although the database is not balanced in terms of gender, the recording conditions and the subject-matter were considered as more influential. Moreover, in [10] no gender differences were detected on two genres of speech – read, and lecture. The total duration of speech material was approximately 3 hours (28% men's; 72% women's), and the total amount of speech is 17,543 spoken words (26% men's; 74% women's). Table 1 summarizes the database characteristics.

Table 1: *The database.*

Speaker	Gender	Speech duration (minutes)	Words count	Number of audio files as input	Date of experiment
SH1	woman	2.14	290	4	4/5/14
BAK	woman	3.96	367	6	4/5/14
BE	woman	19.91	1,809	35	4/5/14
IB	woman	18.96	1,855	39	4/5/14
YV	man	21.72	1,903	40	4/5/14
YBI	woman	24.75	2,272	49	30/5/14
RK	man	24.32	2,639	48	7/4/14
OG	woman	21.19	2,898	43	4/5/14
SH2	woman	27.66	3,510	54	30/5/14
Total		**165**	**17,543**	**318**	

2.2. ASR engines

The two state-of-the-art automatic speech recognition engines that were used are: Google/HTML5 speech recognition system for Hebrew [11] and Nuance Mobile Developer Program – NDEV [12]. Both are closed tools with no possibility to change their acoustic models, and linguistic infrastructure: lexicon (i.e., word list and transcriptions) or language model. Google Voice Search is a free access engine with an Application Program Interface (API). It enables a single query of an audio file (12 seconds long, FLAC format) as an input, and turns back results in JSON format as an output, which is translated into textual format (i.e., transcription). NDEV Mobile is Nuance free product that enables flexible access to their speech models and SR engine. It also has an API interface. In the NDEV engine, the maximal threshold for audio files is of 1 minute length. The audio file required format for both engines is mono *.Wav files, with 16kHz sampling rate, 16bit PCM.

The recorded lectures were first manually cut according to main speech fragments into 318 wav files: The shortest wav file is 4.3 seconds; the longest is 80.9 seconds. The average wav length is 31 seconds with standard deviation of 11.7 seconds. Nevertheless, the engine's output was received as a continuous transcription, according to the original 318 audio input.

The recognition tests were carried during May 2014, and are therefore relevant to the engines' versions at that time. These engines are said to be constantly updated and recognition rates may be changed with each version.

2.3. Performance measures

For each test, the WER is calculated in comparison to an exact transcription reference. WER is derived from Levenshtein Distance Measure that is calculated at the word level and is used to measure the difference between two sequences in information theory. The WER is calculated according to the following formula: Sum of substitutions, deletions, and insertions divided by N, Where: ***S**ubstitution* (S) is a word in the automatic transcriptions that is aligned to a non-identical word in the corresponding manual transcription. ***D**eletion* (D) is a word in the manual transcriptions that is not aligned to any word in the corresponding automatic transcription. ***I**nsertion* (I) is a word in the automatic transcription that is not aligned to any word in the corresponding manual transcription. ***C**orrectness* (C) is a word in the automatic transcription that is aligned to an identical word in the corresponding manual transcription. *N* is the number of words as an input: C+D+S = N.

2.4. Experiment display

An evaluation display was built, which is illustrated in Figure 1. Each audio chunk was named after the speaker (e.g., BAK) and the time span (in seconds) of the chunk (e.g., 12-38). A media player was set in order to be able to edit the exact transcription. The three other transcriptions: Real-Time (RT) Manual, Google's and Nuance's are displayed to the left. The exact transcription is in the right most box. WER components values are also exhibited. At the left most cells there is an option to visualize the 1-best hypothetical alignment according to which the WER was calculated.

Figure 1: *Illustration of the evaluation setting.*

3. Results

The global WER results were as follows: Manual 52% WER; Google 70% WER; and Nuance 49% WER. According to the results demonstrated in Figure 2, the two ASR engines have a stable gap between them and their performance is relatively similar for all the recordings. The manual transcription is spread within a wide range of WERs (31% to 71%). Google's range varies between 59% to 77%, and Nuance demonstrated the minimal range - 43% to 55%.

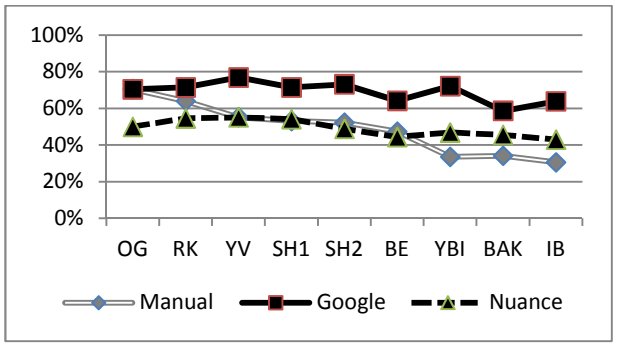

Figure 2: *WERs of two ASR engines compared to manual transcription of eight Hebrew academic lectures (SH1 and SH2 are the same speaker).*

When looking into the WER components – S, D, and I, it is evident that substitution rates are lowest in the manual transcriptions (range between 11%-16.8% of N) with comparison to the two engines. In general, Nuance had fewer S rates (range between 31.9%-40.9%) than Google (range between 37%-54.8%). Deletion rates, on the other hand, were higher for the manual transcriptions (14%-58%) than in the ASRs (maximum of 34%).

3.1. Non-linguistic variables and WER correlation

The speaking rate of each lecturer in terms of words per minutes (WPM) was measured and a Spearman's correlation was conducted. The correlations were not found statistically significant. The highest speaking rate is OG's (136.777 WPM). SH1 also has a rather high rate of 135.335 WPM, yet only at the first part of her lecture, when she welcomed the audience. At the second part of her lecture (SH2) her speaking rate has been lowered to 126.885 WPM. The average speaking rate was 107.585 WPM. These speaking rates are considered as normal to slow rates in Hebrew [13], but future research should use the basic rhythmic units – the syllable and calculate syllables per seconds (SPS) ratios [14]. In further tests, a Pearson correlation was carried to test the correlation between audio lengths and WER values. Although it is explained above that input audio files were automatically reduced to the maximal threshold duration of each of the ASR engines, we wanted to test if this external factor affects WER results. 318 audio files were measured for their duration and no significant correlation was found between the two variables: original audio-length and the WER of Google, Nuance, or the Manual transcriptions. For the same set of 318 files, we also found significant correlation between Google's WER (M=.69, SD=.05) and Nuance WER (M=.49, SD=.04), (R=0.35; $p<0.0001$), and between Nuance WER and the Manual WER (M=.48, SD=.13), (R = .18; $p<0.004$). A T-test showed that Nuance WER was better than Google's (t=15.802, P<.0001); that Manual WER was better than Google's (t=-5.105, P<.001), and that there was no significant difference between Nuance WER and the Manual WER results.

3.2. Linguistic classification of errors

During the ASRs running we have noticed that each engine uses inconsistent writing methods, which affects WER results. The following is a non conclusive classification of linguistic level errors.

3.2.1. Phonetic errors

Although the experiment was carried only on orthographic transcriptions, examination of the words shows that the phoneme recognition for Hebrew works well, while the Language Model is weaker. This is evident when identical or similar phonetic strings are recognized as different words or words sequences. For example, the phonetic sequence [rak**evet** yisrael] 'Israel train' was the 1-best hypothesis for the reference sequence [rak **b**eyisrael] 'only in Israel' (the different phonemes are in bold).

Another phonetic error is the recognition of hesitations disfluencies. In Hebrew these are realized as [e] 'eh' or [em] 'ehm' [15]. An observation on this phenomenon has lead to the conclusion that there is a Hit-and-Clean strategy in the two engines, since hesitations disfluencies are not shown in the output, but sometimes a meaningful word appears instead. For example, a sequence such as *short explanations ehm in Hebrew* was recognize as *short explanations no Hebrew* in one engine and in the other engine the hesitation was ignored and the output was clean: *short explanations in Hebrew*. The fact that the phonetic realization of the most common hesitation disfluencies in Hebrew are [e] or [em] led to irregular use of the lexeme [en] 'no' in one of the engines, which appears in the exact transcription only 31 times, but in Google and Nuance engines' output it appears 133 times and 58 times, respectively. In the exact transcription, the reference, hesitation disfluencies were omitted too, in order not to bias the results. For example, a sequence of four disfluencies [eh] was replaced by one [sheli] 'mine' in one engine and with non-lexeme [ey] in the other. Another lexeme that substituted the disfluencies was [et] 'Accusative marker'.

Last, we argue that phoneme recognition can be seen as a by-product of the WER components ratios. Meaning, higher Sub rates can be interpreted as high phoneme recognition but with lower word recognition. The two engines demonstrate higher Sub ratios with comparison to the manual transcription, while the manual transcription showed higher deletion ratios with

comparison to the two engines. This is not surprising since the stenographer does not record each and every word. A human professional transcription aims to bring the essence of the lecture, and thus, reduces word accuracy for the sake of clear and fluent text.

3.2.2. Textual errors

By far, spelling and varied written methods are the most common errors in both engines, which can also be a reason for the large amount of substitutions. The phenomena can be divided into three categories: Numerals, Full spelling vs. reduced spelling, and (Lack of) punctuation symbols. **Numeral expressions** were found written in several methods with no consistency even in specific types, such as percentage. For example, the phrase "Twenty-five percent" was transcribed "25%" in Nuance and the manual transcriptions, versus the full version "Twenty-five percent" in Google. In other cases there was a hybrid form such as "100 percent" (in Hebrew). Both are correct, but the WER results were carried according to a single *Exact* transcription, which led to discrimination of one of the ASRs. Moreover, after examination of the ASRs' output, a lot of cases were found where a single vocalic letter is the only difference between the reference text (without the vocalic letter) and the automatic output transcription (with the vocalic letter). **Spelling methods**: There are mainly two spelling methods in Hebrew – a full version and a reduced version. The full version uses vowelized letters in a normative form (and hence it disambiguates reading). Again, in cases of discrepancies between the two engines only the normative form was considered as correct. **Punctuations**: The experiment ignored haphazard punctuations. Yet, when there was a correct 1-best hypothesis which contained punctuations (mainly dashes), it was considered a hit.

4. Summary and conclusions

This study analyzed ASR errors with non-linguistic parameters (audio file length, speaking rate) and linguistic parameters (phonetic errors, spelling, etc.), in order to evaluate two state-of-the-art ASRs of Hebrew. We found that the speaking rate and audio file length did not affect ASR's results, but had limited effect on manual RT transcription. As to the WER component analysis, substitution was the main component of WERs, which suggests that acoustic models enable good phoneme recognition. Last, integration of English words and terminology is very common in academic lectures in Hebrew, especially when technological issues are involved, as in our database. Nonetheless, integration of English words and phrases means a complex Language Model. During the last recognition tests we noticed that Google's transcription integrates English phrases, such as '*common sense*', within the Hebrew transcriptions. This advanced feature meets the needs of academic lectures and is a step forward to a natural speech recognizer.

5. Acknowledgements

This research was supported by the Open University of Israel's research fund (grant no. 502532).

6. References

[1] Lippmann, R. P., "Speech recognition by machines and humans", Speech Communication, 22:1-15, 1997.

[2] Googlemobile.blogpost. 2011. Online: http://googlemobile.blogspot.be/2011/12/voice-search-arrives-in-middle-east.html

[3] Biadsy, F., "Google's voice search – focusing on Arabic and Hebrew", Keynote lecture presented in ISCOL conference (June 2013), Ben Gurion University.

[4] Nuance.com. 2012. Online: http://www.nuance.com/company/news-room/press-releases/dragonappsisrael.doc

[5] Wilpon, J. G., Rabiner, L. R., Lee, C. and Goldman, E. R., "Automatic Recognition of Keywords in Unconstrained Speech Using Hidden Markov Models", IEEE Transactions on Acoustic Speech Signal Processing, 38(11):1870-1878, 1990.

[6] Rousseau, A., Deléglise, P. and Estève, Y., "TED-LIUM: an automatic speech recognition dedicated corpus", Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC), 125-129 , 2012.

[7] Biadsy, F., Moreno, P. J. and Jansche, M., "Google's cross-dialect Arabic voice search", IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2012), 4441-4444, 2012.

[8] Pincus, E., Stoyanchev, S., Hirschberg, J.: Exploring features for localized detection of speech recognition errors. In: *Proceedings of the SIGDIAL 2013 Conference*, pp. 132–136. Association for Computational Linguistics, 2013.

[9] Pappu, A., Misu, T., & Gupta, R. (2014). Investigating Critical Speech Recognition Errors in Spoken Short Messages. *Proceedings of 5th International Workshop on Spoken Dialog Systems*, Napa, January 17-20, 2014.

[10] Silber-Varod, V. and Geri, N., "Can automatic speech recognition be satisficing for audio/video search? keyword-focused analysis of Hebrew automatic and manual transcription", *Online Journal of Applied Knowledge Management, 2*(1), 104-121, 2014.

[11] Sampath, S. and Bringert, B., "Speech Input API specification", Google Inc. W3C, 2010. Online: http://lists.w3.org/Archives/Public/public-xg-htmlspeech/2011Feb/att-0020/api-draft.html

[12] Nuance Mobile Developer Program. 2011. HTTP Services for Nuance Mobile Developer Program Clients. Nuance Communications Inc. Online:

[13] Amir, O., & Grinfeld, D. (2011). Articulation rate in childhood and adolescence: Hebrew speakers. Language and speech, 54(2), 225-240.

[14] Finkelstein, M., & Amir, O. (2013). Speaking Rate among Professional Radio Newscasters: Hebrew Speakers. *Studies in Media and Communication*, *1*(1), 131-139.

[15] Silber-Varod, Vered. 2010. Phonological aspects of hesitation disfluencies. *Speech Prosody 2010, The Fifth International Conference on Speech Prosody*, 11–14 May 2010, Chicago, Illinois.