Multi-Dialectical Languages Effect on Speech Recognition

Too Much Choice Can Hurt

Mohamed G. Elfeky Pedro Moreno
Google Inc.
New York, NY, USA
{mgelfeky, pedro}@google.com

Victor Soto
Columbia University
New York, NY, USA
vsoto@cs.columbia.edu

***Abstract*—Research has shown that automatic speech recognition (ASR) performance typically decreases when evaluated on a dialectal variation of the same language that was not used for training its models. Similarly, models simultaneously trained on a group of dialects tend to underperform when compared to dialect-specific models. When trying to decide which dialect-specific model (recognizer) to use to decode an utterance (e.g., a voice search query), possible strategies include automatically detecting the spoken dialect or following the user's language preferences as set in his/her cell phone. In this paper, we observe that user's voice search queries are usually directed to a dialect-specific recognizer that does not match the user's current location, and present a study that shows that automatically selecting the recognizer based on the user's geographical location helps improve the user experience.**

Keywords—multi-dialectical languages; speech recognition; voice search

I. Introduction

Dialects are defined as variations of the same language, specific to geographical regions or social groups. For example, Mexican, Argentine and Castilian are all different Spanish dialects. Similarly, Egyptian, Gulf, Levantine and Maghrebi are the main four Arabic dialects. Multi-dialectical languages, e.g., Arabic, Spanish and certainly English, pose a challenging question to speech recognition systems. Should we build a single recognizer that understands all dialects, or build a different recognizer per dialect? Although dialects of the same language share many similarities, they are often differentiated at several linguistic levels; amongst others: phonological (i.e., how it sounds), grammatical, orthographic (e.g., "center" vs. "centre") and very often different vocabularies. Dialectical groups can also be characterized by the level of intelligibility between its speakers. While dialectical Spanish speakers can generally understand each other without much difficulty, Arabic speakers are often unintelligible or hard to understand each other. It is because of these divergences that computational tools trained or tuned for one specific dialect will break or underperform when tested on another dialect from the same dialectal group. Similarly, global tools simultaneously trained on many dialects fail to generalize well for any of them.

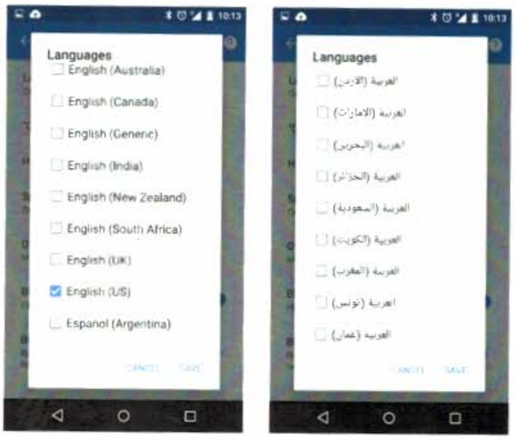
Therefore, state-of-the-art speech recognition systems, including that of Google, have answered that challenging question above by building a different recognizer per dialect. That is, each recognizer is trained both acoustically and linguistically on dialect-specific data. This decision was based on linguistic facts as well as rigorous cross-dialect experimental analysis (e.g., [1]). Google speech recognition system has four recognizers for Arabic, five for Spanish, and eight for English (including three in the works). Table I shows the up to date cross-dialect evaluation for the four main Arabic dialects, using supervised data sets collected from the specific regions. It clearly shows that each dialect-specific recognizer has the best performance over its dialect test set. That is, for example, the Egyptian Arabic recognizer beats the other Arabic recognizers when operating over Egyptian Arabic test sets. Note that each test set comprises manually curated transcribed anonymized utterances that belong to that dialect.

TABLE I. Arabic Cross-Dialect Analysis (Word Error Rate %)

	Recognizer			
Test Set	***Egyptian***	***Gulf***	***Levantine***	***Maghrebi***
Egyptian	**34.0**	39.0	47.8	51.1
Gulf	28.2	**22.1**	31.2	40.0
Levantine	31.2	26.3	**25.7**	39.9
Maghrebi	46.2	41.6	47.5	**26.3**

With that decision, the issue comes now to how to choose which dialect-specific recognizer to use for speech requests from that multi-dialectical language. In the past few years, there has been an emergent effort in the Speech Recognition community to develop automatic dialect identification tools to be later integrated into the ASR pipeline. Several approaches have been proposed to build dialect classifiers. Phonotactic approaches [2, 3, 4] exploit the hypothesis that dialects differ in their phone sequence distributions. Work has also been carried out in the use of prosodic information for dialect identification. For example, in [5] intonational cues are used to discriminate between two German dialects, and in [6] rhythmic differences are used to discriminate between Arabic dialects. Biadsy [7] later showed that in addition to phonotactic features, intonation and rhythmic features help improve Arabic dialect identification [8]. In [9], it was shown that vocalic and consonantal interval length are discriminative features of Arabic dialects and were later used in [10] for dialect identification. However, most of this research is language-dependent and is yet to be found to work in large-scale multi-language speech recognition systems. Moreover, large-scale state-of-the-art automatic language identification [11, 12] systems do not perform well identifying dialects due to the scarcity of acoustic differences between dialects.

Therefore, the decision of which dialect-specific recognizer to use is indirectly handed to the user. Simply, the user selects which language / country he speaks and his voice queries will be directed to a corresponding recognizer. For example, if a user selects English (UK) from the list of language / country pairs available in voice languages settings (Fig. 1), then his/her voice queries will be served by the UK English recognizer. Similarly, the Gulf Arabic recognizer will serve users who select Arabic (Saudi Arabia) or Arabic (Qatar), etc.; and the Latin Spanish recognizer will serve users who select Spanish (Chile); etc. Although giving the user this control sounds ideal, we present here empirical evidence that this does not yield to good user experience.



Languages

□ English (Australia)
□ English (Canada)
□ English (Generic)
□ English (India)
□ English (New Zealand)
□ English (South Africa)
□ English (UK)
☑ English (US)
□ Espanol (Argentina)

CANCEL SAVE

Languages

□ العربية (الأردن)
□ العربية (الإمارات)
□ العربية (البحرين)
□ العربية (الجزائر)
□ العربية (السعودية)
□ العربية (الكويت)
□ العربية (المغرب)
□ العربية (تونس)
□ العربية (عمان)

CANCEL SAVE

Fig. 1. Voice input languages settings screen

II. Claim and Preliminary Analysis

We claim that

The country selection in Google voice input list of language / country pairs does not yield a good user experience.

It is obvious that language / country selection in the phone voice languages settings is not user friendly. Aside from being such a long list of language / country pairs, we believe that having multiple entries for one language might also confuse the speakers of this language. Users come to that screen (Fig. 1) to set the language they speak to enable voice input, and hence are not expecting that the country selection would matter. Therefore, they might end up choosing an arbitrary entry that has this language. Here are the observations that heavily support our claim:

- Almost 60% of the Arabic speech requests originating from Egypt are served by the Gulf Arabic recognizer. This suggests that the users who sent those requests, for some reason, have selected a different Arabic than Arabic (Egypt) as their Google voice input language.
- Only 20% of the speech requests that are served by the Egyptian Arabic recognizer originate from Egypt. This suggests that, for some reason, a significant number of non Egyptian Arabic speakers have selected Arabic (Egypt) as their Google voice input language. This is more plausible than assuming that there are more Egyptians worldwide than in Egypt.
- Almost 70% of the Spanish speech requests originating from Mexico are served by a Spanish recognizer other than the Mexican one.
- More than 20% of the speech requests that are served by the Spain Spanish (Spanish as spoken in Spain) recognizer originate from Latin America.
- Similar numbers were observed for Australian English, Argentine Spanish, etc.

To better visualize the above observations, Fig. 2 shows a heat map for the Arabic traffic over one month, where the colors red, blue, green and black are used to represent Egyptian, Gulf, Levantine, and Maghrebi Arabic traffic, respectively. The other colors in the map are simply produced when mixing two or more of the main colors due to mixed traffic. As expected, the Gulf Arabic traffic (blue) dominates the Gulf region, and the Levantine Arabic (green) dominates the Levant region. However, the Egyptian Arabic (red) does not dominate Egypt where the high percentage of the Gulf Arabic traffic results in that "purple" color produced by mixing red and blue. Also, observe how Egyptian Arabic dominates Yemen rather than the expected Gulf Arabic, and how Maghrebi Arabic (black) is almost non-existent due to perhaps being recently launched.

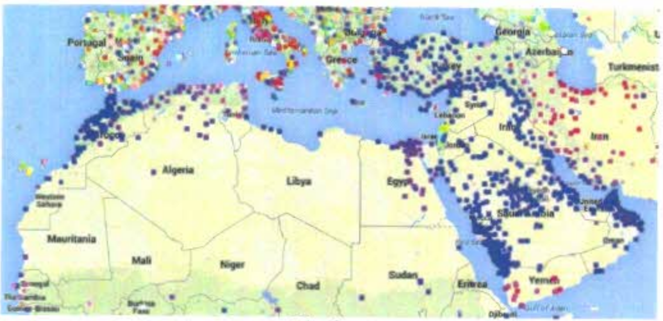


Fig. 2.

Arabic traffic heat mapWe could think of, but later rejected, two counter arguments: (i) users who sent those speech requests have deliberately selected the other dialect since it performs better and provides more accurate results; (ii) those users are travelers and expats. We reject the first counter argument based on the results of cross-dialect manually-rated side-by-side (SxS) experiments presented in Section III. They clearly show that the recognizer that is designed specifically for the dialect spoken in one country outperforms the other dialects' recognizers. We reject the second counter argument not only because the percentage is too high to represent travelers and expats, but also based on having those SxS experiments manually rated by natives of the other country. For example, the SxS experiments analyzing the Arabic speech requests originating from Egypt but served by the Gulf Arabic recognizer will compare the Egyptian recognizer to the Gulf recognizer and will be manually rated by both Egyptian and Gulf raters. If both sets of raters prefer the Egyptian recognizer, as will be shown next, this clearly supports our claim and rejects both counter arguments.

It is perhaps hard to imagine this high percentage of confused users while selecting the voice input language. One explanation could be that those users have never changed the default language set by the manufacturer. Most of the phones sold in Egypt are imported from UAE, and hence the default language/country selection is set to Arabic (UAE), which means the Gulf recognizer. This also might explain Mexico phones having their default as Spanish (US). One very interesting observation that favors this explanation is: almost 99% of the speech requests that are served by the Afrikaans recognizer do not originate from South Africa. The top 3 countries that account for more than 40% of the Afrikaans traffic are Brazil, USA, and India. It is hard to imagine that there are more speakers of Afrikaans in these countries than in South Africa. Knowing that Afrikaans (South Africa) is the first entry in the voice input language settings solves that mystery, and certainly supports our claim.

Whatever the explanation is, there is indeed a bad user experience in terms of bad voice recognition associated with user's voice input language selection, and we would rather have to take control to give users better voice recognition for their speech requests.

III. Extensive Analysis

A. *SxS Experiments*

The first set of experiments we conducted to support our above claim are the side-by-side (SxS) experiments. The objective here is to compare the speech recognizer of the user's selection to a recognizer based on their country. The Google internal SxS framework allows us to select a number of live anonymized speech requests (utterances) and their current transcriptions (speech recognizer output), transcribe them by a different recognizer, and present the results to raters. Each utterance and its two transcriptions are presented to at least three different raters who listen to the utterance and choose which transcription is better (or if both are equally bad). The utterances are anonymized in the sense that all personally identifiable information are filtered. For our purpose, the selection criteria of the live speech requests are the originating country of the utterances and the recognizer that was used to transcribe it. Remember that this recognizer maps directly to the user selection from the voice input language settings. For fair analysis, we do not have control on who the raters are but only where they are from.

Fig. 3 shows the combined results of these experiments. We have conducted 12 of such experiments. Each experiment corresponds to a pair of originating country, and the original recognizer. The recognizer we choose to compare against is the one that corresponds to the same language and the dialect spoken in this originating country. For example, the first line corresponds to the experiment conducted on speech requests originating from Egypt where the original recognizer was Gulf Arabic (almost 60% as observed before). Clearly, we chose the Egyptian Arabic recognizer to compare against. The raters chosen for this experiment were from both Egypt and Saudi Arabia (a Gulf country). Similarly, we conducted two experiments for traffic originating from Argentina: one when the original recognizer is US Spanish, and one when it is Spain Spanish (both account for almost 70% of the traffic). We compare both against the Argentine Spanish recognizer with raters from both Argentina and USA for the former, and Argentina and Spain for the latter. Fig. 3 is a bar chart, where the total length of each bar represents the total percentage traffic. Each bar is divided into four areas: Unchanged represents those utterances which the two recognizers generated the same transcription; Neutral represents those which the raters did not prefer a transcription over the other; Positive represents those which the raters preferred the transcription generated from our chosen country-specific recognizer; and Negative represents those which the raters preferred the original transcription. Except in three cases, Fig. 3 shows clearly the Positives are more than Negatives, which means that the country-specific recognizer outperforms the original one (corresponding to user's selection). This fully supports our claim that the user's selection does not yield good experience. It also suggests that a simple country-based automatic selection of the recognizer outperforms the user's selection.

The three cases when the original recognizer outperformed the country-specific one are studied further. In South Africa, we found that the South African English recognizer has a very high word error rate itself, and hence it is not expected to outperform the US or the UK English recognizers. For Colombia (and the other Latin American countries), we found that the Latin American Spanish recognizer is almost useless. A cross-dialect analysis of the five Spanish recognizers, shown in Table II, revealed that the Mexican Spanish recognizer outperforms the Latin American Spanish recognizer on Latin American test sets. That led us to further decide to get rid of the Latin American Spanish recognizer and use the Mexican one in all Latin American countries except for Argentina. That decision was also supported by a SxS experiment (Table III) comparing both recognizers over live speech requests originating from the affected countries.

TABLE II. Spanish Cross-Dialect Analysis (Word Error Rate %)

	Recognizer				
Test Set	***Latin***	***Argentine***	***Spain***	***Mexican***	***US***
Latin	25.5	27.3	24.5	**23.0**	24.4
Argentine	32.1	**27.0**	29.6	28.9	30.1
Spain	23.8	23.3	**13.3**	19.1	20.9
Mexican	18.3	19.1	16.7	**10.6**	13.7
US	20.0	20.8	17.7	15.3	**12.3**

TABLE III. Latin vs. Mexican Spanish SxS

Sample Size	***Wins***	***Losses***	***Neutral***
500	79	54	367

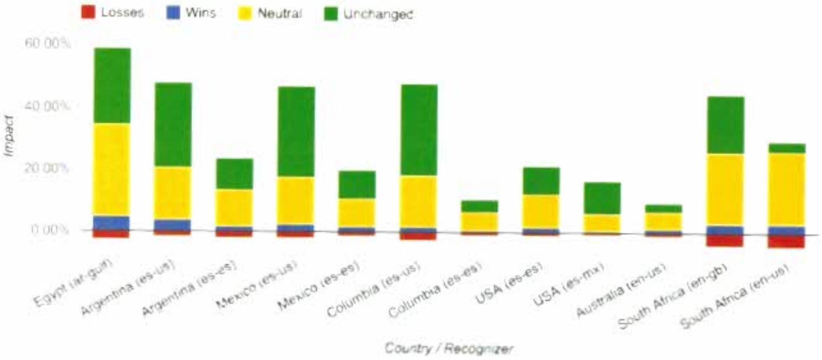


Fig. 3. SxS Impact Analysis

B. Live Experiments

The ultimate goal of ASR systems is to provide the correct transcriptions so that when used in voice search, for example, the search results accurately represent what the user is asking for. The next set of experiments are live experiments to measure how ignoring user's selection corresponding speech recognizer and directing the voice search queries to a country-based recognizer would improve the search experience. Table IV shows four of such experiments in different regions of the world where we randomly directs 10% of the voice search requests originating in that region to a country-based speech recognizer. The metric used here is the user interactions with the search results, which clearly represents the user satisfaction with the search results. Table IV shows that user interactions increase when we use country-based speech recognizers.

TABLE IV. USER INTERACTION RATE (RELATIVE DIFFERENCE %)

Egypt	*Australia*	*Spanish Speaking Countries*	*Worldwide*
+3.67	+0.80	+0.59	+0.06

C. Travelers and Expats Analysis

In order to counter the last counter-argument aforementioned, we further studied how often our anonymized users travel, or more accurately how often they send speech requests from a country different than the one they send most speech requests from. In this analysis, we used the speech requests of users who opted in keeping their audio history, and we only aggregated the collected data without access to any specific user. Here are our findings:

- Around 95% of our users send speech requests from only one country, i.e., they either do not travel or do not use voice input while traveling.
- For the remaining 5%, the distribution of the countries histogram looks like a Zipfian distribution, as shown in Fig. 4a. Almost 80% travel to only one other country, etc.
- Fig. 4b shows the percentage of those 5% "traveling" users versus how much traffic they generate when they travel. For example, 50% of those users generate less than 5% of their traffic while traveling, 75% of them generate less than 20%, etc.

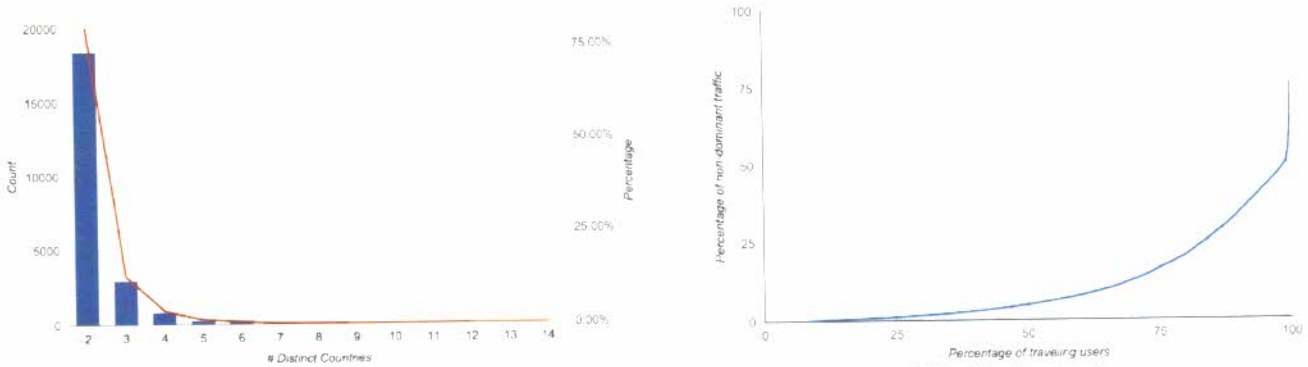
These findings show that only a small percentage of our users use voice input when they travel and when they do, they use it less frequently that they do from their home countries. This proves that the observations we discussed in Section II are clearly not due to travelers or expats, but rather due to the wrong selection in voice input language settings.

CONCLUSIONS

We conclude that although having a different speech recognizer for each spoken dialect of multi-dialectical languages makes sense linguistically, giving the user the control of which speech recognizer to choose counterparts that benefit. Empirical analysis and extensive experiments support our finding, and prove that country-based automatic selection of the speech recognizer outperforms user's selection. Moreover, our experiments have led us to get rid of one Spanish dialect recognizer, which clearly reduces the overall speech recognition system footprint. For future work, we propose to reduce the size of the voice input language setting list by having only one entry for the language. That is, the user selects Arabic, and internally we decide which dialect recognizer to use. This will entail research in dialect identification, acoustic model adaptation, and / or ensemble learning.

ACKNOWLEDGMENT

The authors would like to thank Jeremy O'Brien for contributing the cross-dialect analysis of the five Spanish speech recognizers, and Daan van Esch for contributing the heat map tool.



(a) Histogram of traffic

(b) Traffic when traveling

Fig. 4. Traveling Analysis

REFERENCES

[1] F. Biadsy, P. Moreno, and M. Jansche, "Google's cross-dialect Arabic voice search," ICASSP 2012, pp. 4441-4444.

[2] M. Zissman et al., "Automatic dialect identification of extemporaneous conversational, Latin American Spanish speech," ICASSP 1996, pp. 777-780.

[3] W. Shen, N. Chen, and D. Reynolds, "Dialect recognition using adapted phonetic models," INTERSPEECH 2008, pp. 763-766.

[4] F. Biadsy, J. Hirschberg, and N. Habash, "Spoken Arabic Dialect Identification Using Phonotactic Modeling," EACL 2009 Workshop on Computational Approaches to Semitic Languages.

[5] J. Peters et al., "Identification of Regional Varieties by Intonational Cues An Experimental Study on Hamburg and Berlin German," Language and Speech 45.2 (2002), pp. 115-138.

[6] R. Hamdi et al., "Speech Timing and Rhythmic structure in Arabic dialects: a comparison of two approaches," INTERSPEECH 2004.

[7] F. Biadsy, "Automatic Dialect and Accent Recognition and Its Application to Speech Recognition," PhD Thesis, Columbia U., 2011.

[8] F. Biadsy, and J. Hirschberg, "Using Prosody and Phonotactics in Arabic Dialect Identification," INTERSPEECH 2009.

[9] S. Ghazali, R. Hamdi, and M. Barkat, "Speech Rhythm Variation in Arabic Dialects," International Conference on Speech Prosody, 2002.

[10] M. Belgacem, G. Antoniadis, and L. Besacier, "Automatic Identification of Arabic Dialects," LREC 2010.

[11] I. Lopez-Moreno, J. Gonzalez-Dominguez, O. Plchot, D. Martinez, J. Gonzalez-Rodriguez, and P. Moreno, "Automatic language identification using deep neural networks," ICASSP 2014, pp. 5337-5341.

[12] J. Gonzalez-Dominguez, I. Lopez-Moreno, H. Sak, J. Gonzalez-Rodriguez, and P. Moreno, "Automatic language identification using long short-term memory recurrent neural networks," INTERSPEECH 2014, pp. 2155-2159.