Evaluating Repetitions, or how to Improve your Multilingual ASR System by doing Nothing

Marijn Schraagen, Gerrit Bloothooft
UiL-OTS, Utrecht University, The Netherlands
{M.P.Schraagen, G.Bloothooft}@uu.nl

Abstract

Repetition is a common concept in human communication. This paper investigates possible benefits of repetition for automatic speech recognition under controlled conditions. Testing is performed on the newly created Autonomata TOO speech corpus, consisting of multilingual names for Points-Of-Interest as spoken by both native and non-native speakers. During corpus recording, ASR was being performed under baseline conditions using a Nuance Vocon 3200 system. On failed recognition, additional attempts for the same utterances were added to the corpus. Substantial improvements in recognition results are shown for all categories of speakers and utterances, even if speakers did not noticeably alter their previously misrecognized pronunciation. A categorization is proposed for various types of differences between utterance realisations. The number of attempts, the pronunciation of an utterance over multiple attempts compared to both previous attempts and reference pronunciation is analyzed for difference type and frequency. Variables such as the native language of the speaker and the languages in the lexicon are taken into account. Possible implications for ASR research are discussed.

1. Introduction

Systems for automatic speech recognition (ASR) are challenged by non-native speakers and multilingual or non-standard vocabulary, such as proper names. Non-native speakers have difficulties in choosing the right phonemes, to pronounce them correctly, and to speak fluently. Additional problems arise, also for native speakers and the ASR system itself, when phrases contain words from more than one language, or archaically spelled words. The latter poses serious difficulties for the grapheme-to-phoneme (G2P) conversion, both for the speaker and for the ASR system. When the interaction between speaker and system fails, a speaker can make a new attempt to pronounce the required utterance. Such a repetition may be a semantic variant (rephrasing) or an attempt to improve pronunciation. This strategy is likely to occur for infrequent or multilingual terms. Since repetition is important in human communication, we were interested whether and how repetition may improve recognition results in ASR as well.

Native infrequent terms and spelling, multilinguality, and repetition provide the setting for the Dutch-Belgian Autonomata TOO project (CLST Nijmegen, ELIS Gent, UiL-OTS Utrecht, Teleatlas, Nuance). The general aim of the project is to improve automatic speech recognition for native non-standard and multilingual terms. A specific research goal is to analyse differences in linguistic and phonetic realisation between speakers with various mother tongues, to adapt current G2P models to fit these realisations, and to investigate properties and possible use of repetition in ASR. An orthographically and phonetically transcribed speech corpus has been developed to perform experimental validation.

2. Corpus and speakers

The lexicon used in Autonomata TOO contains names of commercial Points of Interest (POIs) in The Netherlands and Belgium, such as restaurants, hotels, and rental companies. This lexicon contains many infrequent (and therefore lesser known) names and standard nouns, many of them having archaic or otherwise non-standard spelling. Besides this, POIs exhibit a high degree of foreign influence. Our lexicon design included Dutch, foreign (English or French) and mixed (Dutch-English or Dutch-French) names of POIs.

Speech is recorded from native speakers of Dutch, foreign speakers with a linguistically and culturally related background (French and English), and foreign speakers

POI and speaker distribution		Dutch	French/English	Mixed	Total
	Distinct POIs	*120*	*600*	*80*	*800*
Mother tongue	Speakers	Total number of POIs recorded			
Dutch	4*10	1200	6000	800	8000
French/English	20	2400	0	1600	4000
Turkish/Moroccan	20	2400	0	1600	4000
Total	80	6000	6000	4000	16000

Table 1: Autonomata TOO corpus design. Dutch speakers are divided into 4 groups, each recording ¼ of all 800 POI names. All foreign speakers have recorded all of the 200 Dutch and mixed POIs, but no French/English POIs. This leads to 200 recorded POIs for each participant. Every French/English POI is recorded 10 times (by 10 out of 40 Dutch speakers), and every Dutch/mixed POI is recorded 50 times (by 10 out of 40 Dutch speakers and all 40 foreign speakers).

from linguistically and culturally less related immigrant countries (Turkish and Moroccan). With this set-up, we can evaluate the full spectrum of lexicon-related and multilingual issues in a systematic and controlled manner.

Sound recordings were made for all combinations of POI language and speaker background, except for the combination of foreign names spoken by foreign speakers, which is not a priority of the project. Every POI name is recorded for at least 10 subjects (see Table 1).

The group of native Dutch speakers is balanced for gender and age, with two age groups separated by the age of 40. Their birth region is roughly evenly distributed across The Netherlands and Flanders (the Dutch-speaking part of Belgium). These speakers were contacted through university recruitment web sites and contacts from members of the research team. For Turkish and Moroccan immigrant groups the main recruitment criterion for this group was language proficiency. High levels of proficiency, resulting in near-native speech, are less suited for the project, because speech recognition software is already capable of handling such minor accents perfectly well. Therefore, the preferred level of proficiency was around CEF level B2. However, language proficiency is directly linked to the level of participation in society. We had to put in considerable (social) effort to recruit subjects.

3. Corpus recording

A novel, application oriented approach has been used for the recording of the corpus. During the recording session, every utterance was immediately processed by the Nuance Vocon 3200 speech recognition system. The recognizer was set to baseline conditions, using a Dutch standard G2P, Dutch acoustic models, and a 16000 item POI database as recognition lexicon. The speaker could see the recognition result, and if the utterance was recognized incorrectly, the speaker could choose to repeat the utterance. This process continued until either the utterance was recognized correctly or the speaker decided to give up and proceed to the next utterance. All recordings, including the repetitions, were orthographically and phonetically transcribed, while the orthographic transcriptions were tagged for word or syllable insertions, deletions and substitutions.

ASR research based on actual human-computer interactions (including repeated utterances) is often embedded in a dialogue context. Resulting recordings are usually of a less controlled nature, which complicates systematic research on detailed user-system interaction (such as repetition) at the level of correct recognition of individual words or phrases. The Autonomata TOO corpus is designed to facilitate such research.

Although our approach resulted in a systematically recorded corpus, there were some minor deficiencies. Because the speaker controlled the recording interface, it was unavoidable that sometimes the recording protocol was not obeyed in full. This resulted in a few missing recordings. Also, the speech recognition silence detection could abort a recording too early, which could lead the user to abandon the item without completing a full recording. Occasionally an incomplete utterance was recognized correctly, and the system proceeded to the next item. However, the number of incomplete or missing recordings is small (<1%).

4. Repetitions and recognition accuracy

We first analysed the improvement of recognition accuracy for repeated utterances[1]. An improvement was found in all speaker groups. The biggest effect is present in cases where the unfamiliarity with the lexicon and pronunciation is large, i.e. the foreign speakers (see Table 2).

The relative recognition improvement through repetition up to 57 % is surprising, given that the ASR system operated under baseline conditions during the recordings. This means that no G2P modifications have been made to account for irregularities of Dutch POI names as compared to standard language. Moreover, Dutch G2P rules are also applied to French and English POI names which accounts for the high percentage of errors on English and French utterances for Dutch speakers. Regarding acoustic modelling, the Dutch models completely ignore any multilinguality issues considering phoneme realisation. During the repetition procedure the recogniser could not use specific information

mother tongue speaker	**utterance language**	**% errors at 1st attempt**	**% errors at nth attempt**	**improvement (pct point)**	**relative improvement (%)**
Dutch	Dutch	8.1	4.4	+3.7	45.7
Dutch	English/French	26.3	16.6	+9.7	36.9
English/French	Dutch	12.3	5.2	+7.1	57.7
Turkish/Moroccan	Dutch	20.0	9.1	+10.9	54.5

Table 2: Recognition errors using one and *n* attempts (n>=1).

[1] The corpus is still under development. Currently 85% of all speakers has been recorded and transcribed; all presented results are based on this speaker group.

from previous attempts. The baseline set-up included a general speaker adaptation method, but experiments without speaker adaptation show similar relative improvements. All of the above considerations imply that the significant recognition improvement is entirely due to adapted and improved pronunciation by the listener.

The distribution of *n* (number of attempts) is shown in Figure 1. Most successful repetitions are already reached on the second attempt. Using more than 3 attempts generally does not lead to correct recognition.

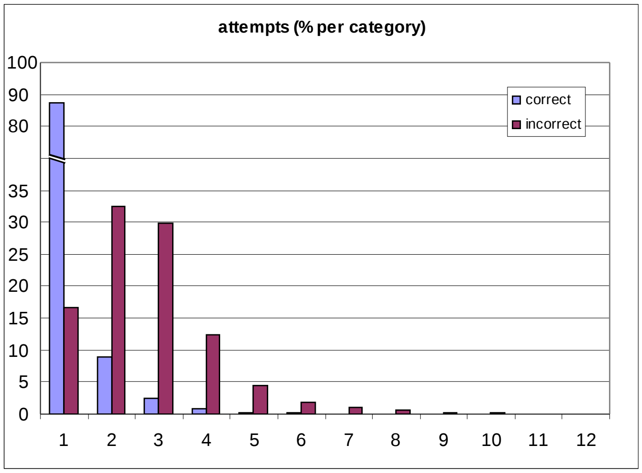

Figure 1: Distribution of repetitions. Percentages are relative to the total number of test lexicon items eventually recognized correctly (left column) or incorrectly (right column).

5. The relative importance of repetition types

On the basis of the phonetic and orthographic transcriptions, repetitions can be divided into four different categories (see Table 3 for examples):

1. Same phonemes, in a number of cases with improved articulation, but often also without any (noticeable) realisation difference.
2. Adapted phoneme realisation (substitution) within vowels, the velar and uvular plosives and nasals, and the voiced fricatives /v/ and /w/, which constitute the majority of phoneme realisation errors in a multilingual setting around Dutch.
3. Structural improvements by insertions or deletions of syllables or phonemes, or by phoneme substitution across broad phonetic categories.
4. Correction of reading errors.

The difference between category 3 (structural variation) and 4 (reading error) is partly a matter of degree. When for example a phoneme is inserted, this could be considered a reading error as well as a structural variant. In the present categorisation, a distinction is made between ‘sloppy reading’ and essential utterance alterations. In Table 3, insertion of /t/ in ‘Verhaghe’ is considered sloppy (and therefore a structural variant), while deletion of a full syllable in ‘Padjelanta’ is considered essentially a reading error. In the majority of cases this distinction is clear.

Figure 2 shows the proportions of applied strategies in the 2^{nd} or further attempts. These repeated utterances are divided in two groups, each with their respective category distribution: repetitions leading to correct recognition and repetitions where recognition still fails (although pronunciation itself might be improved). The strategy of using the same (most likely correct) phonemes as in the previous attempt accounts for 42% of all successful repetitions on average (Figure 2 left column). However, when either the speaker or the G2P system is clearly wrong, exact repetition will not help. This is visible from the large portion of unsuccessful same-phoneme repetitions (Figure 2 right column). The three other strategies, implying a more radical improvement of the realisation, are about equally frequent (around 20% each).

For English and French speakers, the four categories were more evenly used, while for Moroccan and Turkish speakers the repair of reading errors was more prominently needed (see Figure 3). Details for utterance origin (Figure 4) show that in mixed POIs there is more need for structural changes or repair of reading errors.

Repetition category	**Orthography**	**Realisation at attempt *n-1***	**Realisation at attempt *n***
1. Same phonemes	Martha	`'mAr.ta`	`'mAr.ta`
	Asian Tower	`A.zi.An_'tA&u.w$r`	`A.zi.An_'tA&u.w$r`
2. Phoneme adaptation	Huize Orphee	`hA&u.z$_Or.'fe`	`h^&y.z$_Or.'fe`
	Fewaplan	`'fy.v$.plAn`	`'fe.w$.plAn`
3. Structural adaptation	Broeder Jules	`bru.d$r_'jy.lEz`	`bry.d$r_'djylz`
	Verhaghe et Fils	`vEr.haG.t$_e_'fils`	`vEr.ha.G$_e_'fils`
4. Reading repair	Maritiem	`ma_ma.ri.'tim`	`ma.ri.'tim`
	Padjelanta	`'pAt.j$.l$`	`pAt.j$.'lAn.ta`

Table 3: Examples of repetition categories (realisations in the LH+ phonetic alphabet)

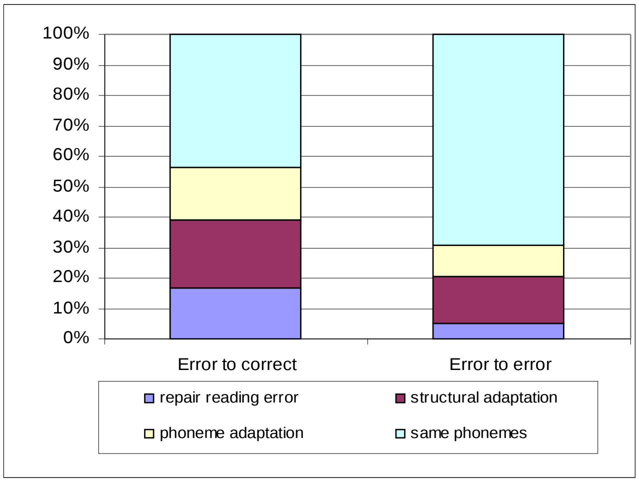

Figure 2: General overview of repetition strategy distribution for both successful and unsuccessful repeated utterances.

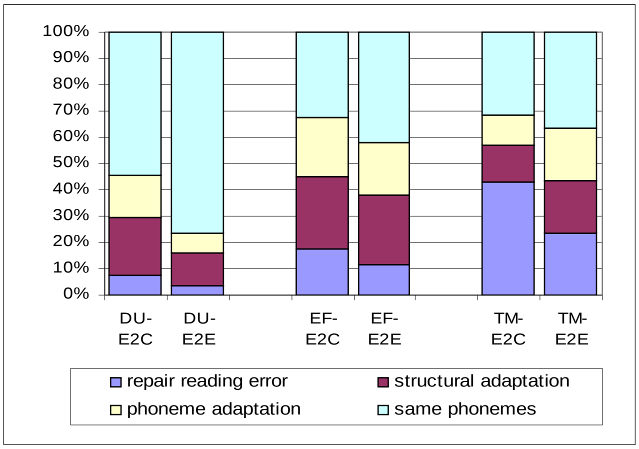

Figure 3: Repetition strategy distribution by speaker origin. DU= Dutch; EF=English and French; TM=Turkish and Moroccan. E2C=Error to correct; E2E=Error to error.

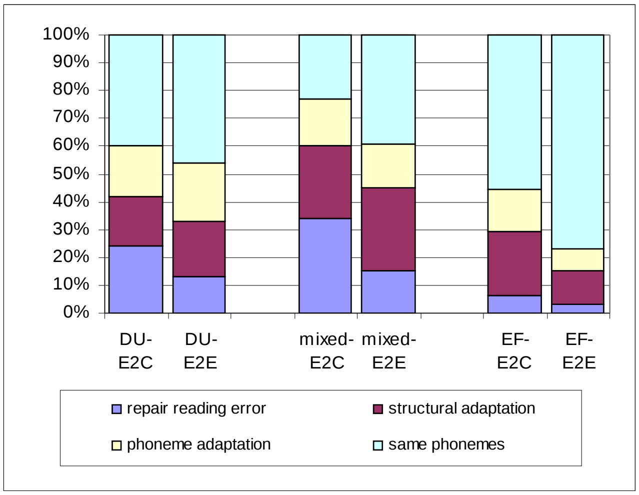

Figure 4: Repetition strategy distribution by utterance (POI name) origin. DU= Dutch; mixed=Dutch-English/Dutch-French; EF=English and French. E2C=Error to correct; E2E=Error to error.

6. Comparison to reference transcriptions

Note that the strategy distributions in figures 2-4 concern transitions between repeated attempts of individual list items. This represents a change the speaker makes from one attempt to the next. However, from these transitions alone we do not know whether the speaker is on the right track, i.e. whether he recognizes his error correctly and adjusts his strategy accordingly, and how the system responds to this behaviour (i.e. if the system performs better when the speaker makes the appropriate adjustments). To address this issue, we have compared the test utterances to a reference phonetic transcription. The reference transcriptions have been created manually by linguists from TeleAtlas, one of the Autonomata TOO project partners.

Figure 5 shows the transition probabilities for the differences between the spoken POIs and the reference transcriptions, for native Dutch/Flemish speakers (left) and Turkish/Moroccan speakers (right). Differences are shown for incorrectly recognized utterances only. It shows, for instance, that incorrectly recognized utterances for Dutch/Flemish speakers were pronounced correctly (EQ) on the first attempt in 38% of the cases, while 18% of the misrecognized utterances initially had a phoneme difference with the reference transcription (PHO) and 52% of the cases showed a structural variant (STR), which dominantly were English and French POIs. Dutch/Flemish speakers hardly made reading errors. For typically pronounced POIs (EQ), 40% of the cases was repeated without success at least once (the self-loop). The repetition was the final attempt (either correct recognition or failure) in 47% of the cases, while for 7% the correct pronounciation was changed into a structural variant. The remaining 6% was either followed by a reading error or a phoneme difference (not shown).

It should be noted that the speaker groups had different tasks. The test set for native speakers consisted for 80% of English and French POIs (see Table 1), which were hard to recognize because of the baseline Dutch G2P. Foreign speakers only read Dutch POIs. In Table 2 recognition results are being specified according to utterance origin, the percentages shown in Figure 5 are weighted averages.

For Turkish and Moroccan speakers, transitions going out of the EQ and PHO categories are rare. This speaker group did use these categories, but the utterances usually were recognized correctly on the first attempt. Recall that these are Dutch POIs recognized using a Dutch G2P. Therefore, pronunciation corresponding to (EQ) or close to (PHO) the reference transcription usually means a close correspondence with the G2P transcription, and correct recognition Because of the low number of utterances involved, percentages going out of the EQ and PHO categories have been omitted here.

In both graphs of figure 5 we can see that there are connections between the four pronunciation categories, indicating the use of a repetition strategy. However, the percentages are relatively small, mostly less than 10%. The exception is the correction of reading errors (ERR) by foreign speakers, with a quite high probability (23%) of turning into a structural variant (STR, which is less severe than a reading error). However, in most cases speakers either stay in the same category (30-40% of the time), or end the current list item. This means that phonetic differences and structural variants are usually not corrected by the speaker. It is possible to change pronunciation within the same category (except of course for the category EQ), but in most cases exactly the same pronunciation is repeated. This corresponds to the high percentage of same phoneme repetitions in previous sections (figures 2-4), where the utterances have been compared to the previous attempt as opposed to comparison to a reference transcription as in the current section.

The graph for native Dutch speakers contains no node for reading errors, meaning that less than 5 percent of all incorrectly recognized utterances resulted in a reading error on the first attempt. Furthermore, the graph for native Dutch speakers shows a relatively high percentage of utterances being repeated (40%) while the observed speech was consistent with the reference transcription. This is due to the portion of foreign words in the test set, which is much higher for native speakers (see section 2). The reference transcription for foreign POI's can be very different from the baseline Dutch G2P transcription, resulting in recognition errors even though a POI is pronounced correctly.

The node labeled ‘end’ contains both successful and failed attempts, in more or less equal proportion. This holds even for reading errors, which can be quite severe (for example leaving out an entire word). The robustness of the recognizer for (even severe) errors complicates the repetition strategy analysis. For native speakers, we observe only a small probability for the transition from PHO (phoneme difference) to EQ (phonemes equal), for example. This can be accounted for by a number of factors: speakers may not know how to correct their phoneme error; or speakers are unwilling to correct their error (and rather accept an incorrect recognition result); but in many cases speakers do not need to correct their error because the utterance is already recognized correctly.

We now return to the questions posed at the start of this section: does the speaker adjust his pronunciation according to his errors, and does the system perform better when the speaker makes the appropriate adjustments? The answer to the first question is: yes, but only to a small degree. However, this is partly due to the robustness of the system for errors. The second question has to be answered negative: no clear correspondence can be found between speaker corrections and recognition performance. Moreover, if an error is made, the category of that error does not seem to be of major importance for the performance of the system.

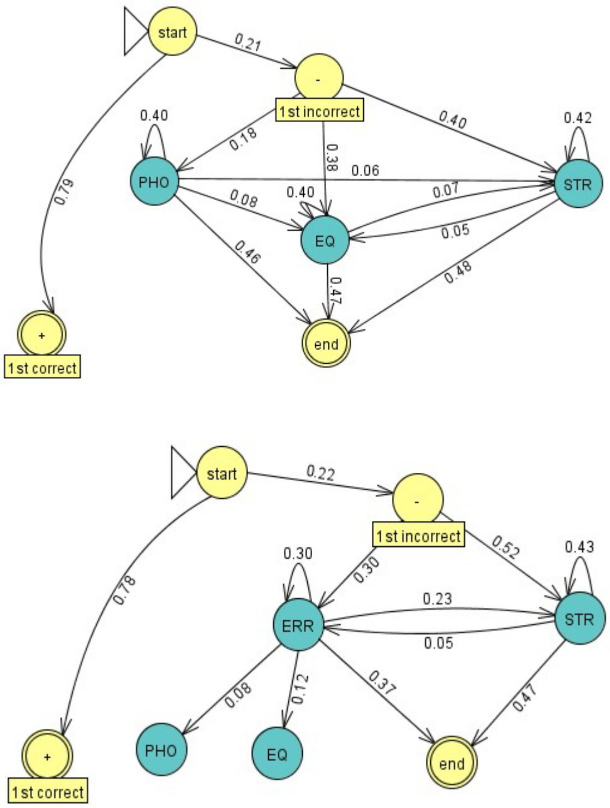

Figure 5: Transition probabilities of repetition strategies for incorrectly recognized utterances (left branches show correct recognition on the first attempt). Top: native speakers of Dutch on Dutch and French/English POIs. Bottom: Turkish/Moroccan speakers on Dutch POIs. PHO=phoneme difference, EQ=same phonemes, ERR=reading error, STR=structural variant, all relative to the reference transcription. Transitions with low probability (<5%) are excluded from the graphs for reasons of clarity of presentation.

7. Conclusion

Speech recognition can be deteriorated by poor reading and speaker pronunciation proficiency. In many cases, it seems that acoustic fine tuning using the same phonemes leads to correct recognition (although such trial and error attempts can be unsuccessful as well). This suggests that the ASR system could benefit from more robust acoustic modelling in the first place. As long as pronunciation errors have systematic phonemic properties, an ASR system could benefit from G2P adaptation and multilingual acoustic modelling[2]. This could help in case of phoneme errors, but these errors constitute only a part of all problems. It is often assumed that this kind of modelling, or *how* something is

[2] See e.g. Van den Heuvel H.; Réveil B. and Martens J.-P., “Pronunciation-based ASR for names”, in Proc. Interspeech, pp 2991-2994, Brighton, UK, 2009.

said (the phonetic realisation category), is the key issue in improving (multilingual) ASR. However, our results seem to indicate that also structural realisation issues and reading errors, or *what* is being said rather than *how*, are at least as important and much more difficult to anticipate by the system.

Our results show that ASR accuracy can be significantly improved without changing the system at all. Just asking the user to repeat what he said, like in everyday human communication, already simplifies the difficult problem of native non-standard and multilingual speech recognition considerably. It is shown that speakers generally do not improve their pronunciation after an initial error, but that the performance of the system can nevertheless benefit from multiple attempts.

8. Acknowledgements

The presented work was carried out in the context of the Autonomata Too research project, granted under the Dutch-Flemish STEVIN program. The corpus will be made available through the TST-Centrale (Dutch-Flemish Language and Speech Technology Repository).