A Minimum Classification Error approach to pronunciation variation modeling of non-native proper names

Line Adde[1], *Bert Réveil*[2], *Jean-Pierre Martens*[2], *Torbjørn Svendsen*[1]

[1]Department of Electronics and Telecommunications, NTNU Trondheim, Norway
[2]ELIS, Ghent University, Belgium
line.adde@iet.ntnu.no

Abstract

In automatic recognition of non-native proper names, it is critical to be able to handle a variety of different pronunciations. Traditionally, this has been solved by including alternative pronunciation variants in the recognition lexicon at the risk of introducing unwanted confusion between different name entries. In this paper we propose a pronunciation variant selection criterion that aims to avoid this risk by basing its decisions on scores which are calculated according to the minimum classification error (MCE) framework. By comparing the error rate before and after a lexicon change, the selection criterion chooses only the candidates that actually decrease the error rate. Selecting pronunciation candidates in this manner substantially reduces both the error rate and the required number of variants per name compared to a probability-based baseline selection method.
Index Terms: speech recognition, proper names, pronunciation modeling, minimum classification error

1. Introduction

One of the most difficult and complex tasks in speech-based form-filling applications is posed by proper names. This is largely related to the fact that a name may be selected from a set of several thousand names and a considerable number of these names are likely to be non-native names. The latter are particularly challenging since they can be pronounced in numerous ways by the users. An individual speaker's pronunciation is likely to be influenced by several sociocultural factors, such as regional background, gender, education and age [1].

Since the pronunciation of a proper name often deviates from what conventional pronunciation rules would predict, the automatic generation of pronunciation variants for these names is a difficult task. Nevertheless, there exist data-driven schemes that can generate acceptable pronunciation variants in specific domains, such as city names or person names, after training on a modest amount of training examples. An example of such a scheme is the grapheme-to-phoneme (g2p) phoneme-to-phoneme (p2p) tandem proposed in [2]. In [3] it is shown that adding the most likely pronunciation variants to the lexicon increases the recognition accuracy. However, we argue that the gain could be further increased by introducing better selection criteria for adding pronunciation variants.

Several previous studies have tried to accomplish this by optimizing different types of criteria, such as frequency of occurrence [4] and acoustic likelihoods [5] of variants in a training set, among others. Unfortunately, there is no direct relationship between these selection criteria and the recognition error rate. Recently however, the authors of [6] proposed to adopt the minimum classification error (MCE) criterion for selecting the most distinctive pronunciation variants.

In this paper we investigate the selection of pronunciation variants on the basis of a criterion that is directly related to the actual recognition error. We further argue that every name should be represented by a set of complementary pronunciation variants, defined as variants correcting different types of recognition errors. We therefore propose an iterative approach in which the lexicon is gradually updated. In every iteration only the candidate variant causing the largest error rate reduction is added for each name. To estimate the number of errors introduced by a particular candidate, MCE scores calculated before and after the addition of the candidate are compared. The proposed approach differs from that of [6] because it calculates the MCE scores on the basis of likelihood scores produced by the recognizer after decoding a name utterance. Furthermore, we take previously selected variants into consideration when selecting new additions to the lexicon, rather than choosing the variants separately.

2. Pronunciation variant selection by error minimization

Traditionally, lexical pronunciation variation modeling first generates a set of candidate variants, each with a probability score, and then retains the most likely ones among them. However, this approach does not necessarily lead to the best recognition performance. In this section we introduce an alternative selection criterion that is directly correlated with the recognition performance.

Let us consider the variant selection problem as a decision problem. Suppose that the set of names is $\mathcal{W} = \{W_1, W_2, \ldots, W_K\}$, and that for some name $W_k \in \mathcal{W}$ we have a set of training utterances $\mathcal{X}_k = \{X_{k1}, X_{k2}, \ldots, X_{kN}\}$ and a set of candidate pronunciation variants $\mathcal{V}_k = \{V_{k1}, V_{k2}, \ldots, V_{kI}\}$. The aim is then to find the variant V_k^* that minimizes the risk of introducing recognition errors. If $\mathcal{L}_k(\mathcal{X}_k; \Lambda_c)$ represents the *expected loss* of recognition accuracy for the training utterances $\mathcal{X}_k$ when using a canonical lexicon and $\mathcal{L}_k(\mathcal{X}_k; \Lambda_{ki})$, the corresponding loss when using the canonical lexicon extended with variant V_{ki}, then

$$V_k^* = \arg\max_{\mathcal{V}_k} \left(\mathcal{L}_k(\mathcal{X}_k; \Lambda_c) - \mathcal{L}_k(\mathcal{X}_k; \Lambda_{ki})\right) \qquad (1)$$

The expected loss of an arbitrary model Λ is usually obtained as the accumulation of contributions $l_k(X_{kn};\Lambda)$ emerging from the available training utterances X_{kn} of name W_k:

$$\mathcal{L}_k(\mathcal{X}_k;\Lambda) = \sum_{n=1}^{N} l_k(X_{kn};\Lambda) \qquad (2)$$

Obviously, $l_k(X_{kn};\Lambda)$ must be a measure of the chance that name utterance X_{kn} is misrecognized by a system using the lexicon denoted by Λ. To define such a measure, we adopt the MCE framework [7]. In that framework one departs from a set of discriminant functions $g_l(X;\Lambda)$ with $l = 1, \ldots, K$. By definition the recognized class $C(X_{kn})$ for utterance X_{kn} is correct if the k-th discriminant function is the largest one:

$$C(X_{kn}) = W_k \quad \text{if} \quad g_k(X_{kn};\Lambda) = \max_l g_l(X_{kn};\Lambda) \tag{3}$$

Most modern speech recognition applications rely on log likelihood scores to make a decision, and consequently, these scores can act as discriminant functions.

If J_{kn} is the set of the most likely name hypotheses proposed by the recognizer for utterance X_{kn}, one can define a *misclassification measure*, $d_k(X_{kn};\Lambda)$ as

$$d_k(X_{kn};\Lambda) = -g_k(X_{kn};\Lambda) + \log\left[\frac{\sum_{j\in J_{kn};j\neq k}^{K} e^{g_j(X_{kn};\Lambda)}}{\text{card}(J_{kn})-1}\right] \tag{4}$$

This measure compares the log likelihood score of the correct hypothesis (k) with the log of the average likelihood of the competing hypotheses.

Finally, in order to map the misclassification measure of (4) to a zero-to-one continuum, the loss function is defined as

$$l_k(X_{kn};\Lambda) = \frac{1}{1+e^{-d_k(X_{kn};\Lambda)}} \tag{5}$$

If the loss is close to zero, it means that the utterance is likely to be correctly recognized using lexicon Λ. The larger the measure is, the larger the risk for an incorrect recognition of the utterance.

3. Experimental set-up

For our experimental study we worked with name utterances selected from the Autonomata Spoken Name corpus (ASNC) [8]. We selected all utterances by native Dutch speakers of 441 unique English names, person names (first name + family name) as well as geographical names (street names and city names). Most names were spoken by six speakers either from Flanders or from the Netherlands, but a couple of names were recorded in both regions, and yielded twelve instead of six utterances. In total, we selected 2760 utterances.

Since our method can only select pronunciation variants of names for which we have training utterances, we can only assess the positive effect of the selected variants on the recognition of these names if the test set also contains utterances of these names. The full data set was therefore divided in a test set comprising one third of the utterances of each unique name and a training set comprising the remaining utterances. The division was made in such a way that there was no overlap in speakers between the two sets.

The recognition engine was the state-of-the-art Nuance VoCon 3200[1] recognizer, running with a standard monolingual acoustic model trained on Dutch speech from Flemish and Dutch speakers. The grammar was a loop of the names included the lexicon.

3.1. Transcription variants

The ASNC corpus is delivered with two phonetic enrichments: a typical Dutch pronunciation (TY) of every name appearing in the name list, and an auditorily verified (AV) transcription of each name utterance. The latter transcription is the best nativized transliteration of what a human expert actually heard when listening to the utterance.

Since we wanted to investigate variant selection methods, we needed to create a pool of candidate variants from which we can choose. To that end we first used the Dutch and English g2p-converters embedded in the Nuance RealSpeak text-to-speech system[2]. Then we trained a Dutch and an English p2p converter [2], departing from the Dutch and the English g2p transcription respectively, to generate variants which approached the auditorily verified transcriptions of the available training utterances. During the variant generation stage, each p2p converter was allowed to generate up to 10 variants per name, but only if their probability exceeded a threshold which was specified as a fraction (we used 0.2) of the probability of the best variant. If the input g2p transcription was not among the created variants, it was added a posteriori with a probability that was equal to the above threshold. This procedure yielded in total 4531 pronunciation variants (and corresponding probabilities) for the 441 unique names.

4. Experimental results

To set the reference, we have conducted recognition tests with lexicons comprising the following transcriptions per name: all auditorily verified transcriptions encountered in the training set (AV), just the typical transcription (TY), the English and Dutch g2p transcriptions pooled together to cover the general pronunciation in both languages (DUNENG g2p) and finally the pool of 4531 transcriptions (DUNENG g2p-p2p) that was created using the procedure described in the previous section.

Table 1 shows for each lexicon the obtained name error rate (NER) as well as the size of the lexicon, defined as the total number of pronunciations it contains. Note that a name is only considered correct if all of its constituents (words) are correct.

Lexicon	Size	NER
AV	1331	4.5%
TY	441	6.1%
DUNENG g2p	876	7.0%
DUNENG g2p-p2p	4531	3.7 %

Table 1: *Number of pronunciation variants in the lexicon and Name Error Rate (NER) for four reference lexicons*

Adding variants to the canonical lexicon DUNENG g2p reduced the NER by almost 50% relative. The variants generated by the g2p-p2p converter even outperformed the AV transcriptions, which speaks in favour of the g2p-p2p approach.

In the tables given in this section, name error rates deemed significantly different from the baseline by the Wilcoxon signed rank test (5% significance level) are marked in bold.

4.1. Baseline selection method

Our baseline method for selecting the variants from the set of 4531 variants was simply to use the variant probabilities generated by the p2p converters as the selection criterion. The results obtained with this method are listed in the left columns of Table 2 and Table 3 as a function of M, the maximum allowed

[1] `http://www.nuance.com/vocon/3200/`

[2] `http://www.nuance.com/realspeak/`

number of variants per name. For each M we also mention the actual size of the corresponding lexicon.

4.2. Selecting variants with an MCE criterion using a single recognition pass

A computationally efficient way of identifying the candidate variants using the MCE framework is the following: recognize all training utterances X_{kn} using a grammar loop of all 4531 candidate pronunciation variants in the lexicon, retain for every utterance the maximally 25 most likely *pronunciation* hypotheses to form the set H_{kn} and calculate the MCE score for every candidate pronunciation variant V_{ki} occurring in H_{kn} using equations (4) and (5). Candidate pronunciations of name W_k not appearing in H_{kn} are given an MCE score of 1 for that utterance. The total MCE score of a candidate pronunciation V_{ki} ($i = 1, \ldots, I$) is then defined as the sum of the MCE scores calculated for each X_{kn} ($n = 1, \ldots, N$) (equation (2)). By taking these MCE scores as the selection criterion, we obtained the results listed in the right columns of Table 2 (as a function of M). When selecting only one or two variants, the MCE method outperformed the baseline method, but when selecting more pronunciations, the situation was reversed. None of these results however, were deemed statistically significant.

	Baseline		MCE single-pass	
M	Size	NER	Size	NER
1	441	8.7%	441	7.5%
2	876	7.1%	876	6.4%
3	1304	6.0%	1304	6.0%
4	1701	5.3%	1701	6.1%
5	2071	5.0%	2071	5.4%
6	2413	4.7%	2413	5.0%
7	2716	4.6%	2716	4.9%
8	2990	4.6%	2990	4.3%
9	3228	4.3%	3228	4.3%

Table 2: *Size and NER of the lexicon created with the baseline and the single-pass MCE variant selection method.*

One reason for this somewhat disappointing result is that two similar pronunciation variants are likely to get similar MCE scores. If the MCE score is low, then both variants will be highly ranked for inclusion in the lexicon although they do not describe complementary pronunciation phenomena.

4.3. Selecting variants with an MCE criterion using M recognition passes

To overcome this problem, we abandoned the single-pass batch approach and switched to a multipass iterative approach. Per iteration, the lexicon emerging from the former iteration was supplemented with one variant of each name, namely the variant that most effectively could reduce the NER attainable with this new lexicon. If no variant could reduce the NER, the lexicon was not changed and no further attempts to add variants for that name were made.

In the first iteration, we generated a lexicon comprising the best variant for each name. To that end we started with an initial lexicon comprising the English g2p transcription of each name. Then, the following procedure was performed for every available candidate pronunciation V_{ki} of name W_k:

1. *replace* the g2p transcription in the initial lexicon by this candidate pronunciation,
2. perform a recognition on all the training utterances X_{kn} of name W_k using this temporary lexicon (with a word grammar containing the 441 names) and collect the maximally 25 most likely *name* hypotheses proposed by the recognizer together with their likelihood scores,
3. calculate the total MCE score of the examined name according to equation (2) given a model Λ_{ki} representing the English g2p transcription of name W_k in the initial lexicon being replaced by V_{ki}.

The variant V_k^* that yielded the lowest total MCE score for name W_k was finally added to the new lexicon.

The procedure in the subsequent iterations ($m = 2, ..., M$) was very similar, but with two differences: (1) the initial lexicon now meant the lexicon emerging from the previous iteration, and (2) for investigating variant V_{ki} of name W_k a new lexicon was created by *adding* V_{ki} to the initial lexicon. The best candidate pronunciation was added if the expected loss was reduced with respect to that of the initial lexicon. This approach ensured that candidate pronunciations were only added if they corrected problems that were left unhandled by the initial lexicon.

The right columns of Table 3 show the results as a function of M again. The figures reveal that the new approach significantly outperformed the baseline method for $M = 2, 3, 4, 5$. Furthermore, it needed to select no more than 3 pronunciations per name to attain a performance that was even surpassing that of an equally large lexicon comprising all the auditorily verified transcriptions found in the training set (see Table 1).

Iterations	Baseline		MCE iter	
M	Size	NER	Size	NER
1	441	8.7%	441	7.1%
2	876	7.1%	876	**4.7**%
3	1304	6.0%	1249	**4.3**%
4	1701	5.3%	1488	**4.1**%
5	2071	5.0%	1625	**4.0**%
6	2413	4.7%	1700	4.1%
7	2716	4.6%	1767	4.1%
8	2990	4.6%	1840	4.2%
9	3228	4.3%	1840	4.2%

Table 3: *Size and NER of the lexicon created with the baseline and the iterative MCE based variant selection method.*

4.4. Testing with a much larger vocabulary

The approach described in the previous section gave a significant NER reduction compared to the baseline method, but nevertheless, it could not beat the method of dumping all 4531 available variants in the lexicon (see Table 1). This is most probably because the vocabulary is too small to expose the increased lexical confusion caused by the inclusion of too many variants. Therefore, we repeated the last recognition experiment, but this time with a vocabulary of 22,441 names: the 441 names appearing in the training and test set and 22k "filler names".

To obtain new reference performances, we created four new reference lexicons in the same way as before. But, since there were no AV and TY transcriptions available for the filler names, the one (case TY) or the four (case AV) most probable g2p-p2p transcriptions were used instead. The results in Table 4 confirm our expectation that the system using all available transcriptions is no longer the best. It is actually only marginally better than the canonical lexicon DUNENG g2p.

Lexicon	Size	NER
AV	88 879	18.8%
TY	22 408	23.5%
DUNENG g2p	44 465	26.2%
DUNENG g2p-p2p	289 258	24.1%

Table 4: *Size and NER of the lexicon of the reference lexicons that ware created for the case of a 22k vocabulary.*

We were hoping that by more carefully selecting the set of variants our iterative method would be able to outperform the canonical lexicon, and even to compete with the AV lexicon.

Since the proposed variant selection method could only be applied on names for which training utterances were available, we conducted experiments in which only the original 441 names got optimized variants. The filler names were left with their M most probable variants according to the Dutch and English p2p converters. Table 5 shows that the baseline (probability-based) selection method was again significantly outperformed by the MCE-based selection method (iterative approach) for all values of M. Moreover, the performance gain was relatively larger than in the case of a small vocabulary.

Iterations	Baseline		MCE iter	
M	Size	NER	Size	NER
1	22 408	27.8%	22 408	**23.3**%
2	44 804	24.8%	44 804	**19.7**%
3	67 142	22.3%	67 085	**19.0**%
4	89 257	21.5%	89 013	**19.8**%

Table 5: *Size and NER of the lexicon for the baseline and the iterative MCE method in case of a 22k vocabulary.*

The best result (for $M = 3$) demonstrates that pronunciation variants can be very helpful to improve the recognition accuracy. There was a relative gain of almost 28% (from 26.2% to 19%). The best performing lexicon was actually reaching the same accuracy as the AV lexicon. The data support our hypothesis that the MCE-approach is more resistant to the risk of increasing the lexical confusability.

5. Conclusions and future work

In this paper we proposed a new pronunciation variation modeling approach that was effective for the recognition of English proper names spoken by non-native English speakers (Dutch speakers in this case). In particular, we proposed a pronunciation variant selection criterion based on the Minimum Classification Error (MCE) framework. The criterion directly took the recognition process into account by using its log likelihood scores as discriminant functions. The iterative nature of the proposed approach ensured that only variants actually decreasing the error rate were included in the final lexicon.

In a small vocabulary test (only 441 names), the iterative approach significantly outperformed a baseline method which selected variants generated by a g2p-p2p converter tandem on the basis of probabilities assigned to these variants by this tandem. For the cases of one and two pronunciation variants per name, the error rate reductions were 18.4% and 33.8% relative. In spite of this, there was no gain with respect to a lexicon holding all pronunciation variants being generated by the g2p-p2p converter. However, in a control experiment with a much larger vocabulary (22,441 names), the full lexicon did not cause any improvement anymore whereas the best lexicon emerging from the newly proposed selection method yielded an improvement of about 28% relative over a canonical lexicon of Dutch and English g2p transcriptions.

One acute problem of the proposed approach is its high computational load due to the required decodings of all training utterances of a name for every pronunciation candidate of that name at every iteration. Therefore, further optimizations of the method are necessary in order to make it suitable for very large vocabularies. Another limitation is the inability to select good pronunciation variants for unseen names. An effort will therefore be made to formulate the variant selection method in terms of more generic mechanisms, e.g. the ones modeled by the p2p converter. Finally, we plan to utilize the pronunciation variants that were found to yield the lowest error rate to generate new improved pronunciation variants.

6. Acknowledgments

The presented work has been funded by the Research Council of Norway as part of the research project *Digitale utfordringer*. The work was conducted in cooperation with ELIS, University of Ghent, in the context of the projects Autonomata Too (granted under the Dutch-Flemish STEVIN program) and TELEX (granted by Flanders FWO).

7. References

[1] Eklund, R. and Lindström, A. (2001), "Xenophones: An investigation of phone set expansion in Swedish and implications for speech recognition and speech synthesis", in Speech Communication, 35 (1-2), 81-102.

[2] Yang, Q., Martens, J.P., Konings, N. and van den Heuvel, H. (2006), "Development of a phoneme-to-phoneme (p2p) converter to improve the grapheme-to-phoneme (g2p) conversion of names", in Proc. LREC, 287-292, Genoa, Italy.

[3] Van den Heuvel, H., Réveil, B. and Martens, J.-P. (2009), "Pronunciation-based ASR for names", in Proc. Interspeech, 2991-2994, Brighton, UK.

[4] Kessens, J.M., Wester, M. and Strik, H. (1999), "Improving the performance of a Dutch CSR by modeling within-word and cross-word pronunciation variation.", in Speech Communication, 29 (2-4), 193-207.

[5] Holter, T. and Svendsen, T. (1999), "Maximum likelihood modelling of pronunciation variation", in Speech Communication, 29 (2-4), 77-191.

[6] Vinyals, O., Deng, L., Yu, D. and Acero, A. (2009), "Discriminative pronunciation learning using phonetic decoder and Minimum-Clasification-Error Criterion", in Proc. ICASSP, 4445-4448, Taipei, Taiwan.

[7] Juang, B.-H. and Chou, W. and Lee, C.-H. (1997), "Minimum Classification Error Rate Methods for Speech Recognition", in IEEE Trans. Speech and Audio Proc., 5(3), 257-265.

[8] Van den Heuvel, H., Martens, J.-P., D'hoore, B., D'hanens, K. and Konings, N. (2008),"The AUTONOMATA Spoken Names Corpus", in Proc. LREC, Marrakech, Morocco.