Automatic evaluation of quantity contrast in non-native Norwegian speech

Ingunn Amdal, Magne H. Johnsen, Eivind Versvik

Department of Electronics and Telecommunications
Norwegian University of Science and Technology (NTNU), Trondheim, Norway
{ingunn.amdal,mhj}@iet.ntnu.no

Abstract

Computer assisted language learning (CAPT) has been shown to be effective for learning non-natives pronunciation details of a new language. No automatic pronunciation evaluation system exists for non-native Norwegian. We present initial experiments on the Norwegian quantity contrast between short and long vowels. A database of native and non-native speakers was recorded for training and test respectively. We have used a set of acoustic-phonetic features and combined them in a classifier based on linear discriminant analysis (LDA). The resulting classification rate was 92.3% compared with a human rating. As expected, vowel duration was the most important feature, whereas vowel spectral content contributed insignificantly. The achieved classification rate is promising with respect to making a useful Norwegian CAPT for quantity.

1. Introduction

Mobility over country borders is increasing. People on the move have various mother tongues, but all have the same need for fast learning of the new native language to be able to participate in the society both socially and professionally. Computer assisted language learning can help people to achieve these skills. It has been shown that such tools are especially effective in the starting phase of learning a new language and for pronunciation training. Computer-assisted pronunciation training (CAPT) systems have a number of important advantages. CAPT is always available and can be used everywhere. CAPT is effective in that it creates a one-to-one teaching situation and is thus ideal for individual learning progress. Further, CAPT allows users to make errors without any loss of prestige, which has a positive effect on the learning process, [1].

Presently there exists no CAPT system for Norwegian, and research results are limited. This paper reports initial experiments on an automatic pronunciation evaluation module of a Norwegian CAPT system. Experience from studies on other languages have been used when suitable, e.g. [1] (Swedish), [2] (English), [3] (French), and [4] (Dutch).

When designing a CAPT system it is important to select pronunciation exercises that are beneficial for the users. Many pronunciation properties that are crucial in Norwegian (L2) may not be important in the user's mother tongue (L1). These properties are obvious candidates for CAPT training. Intonation and duration are examples of such properties. Native perception of these two properties in non-native Norwegian speech is investigated in [5]. In this paper we focus on a specific duration property; the contrast between long and short vowels. This property is phonemic in Norwegian, but not important in many L1s.

The users need corrective feedback about pronunciation errors. It is important to do this in a form that is understandable to a wide variety of users. A CAPT system thus needs an automatic evaluation of the pronunciation, often using techniques from spoken language technology. High level of detail of the automatic pronunciation evaluation is needed to give sufficient feedback. This means information at the phone level or at even more detailed articulatory level. In addition, some measure of the deviation from a native pronunciation is useful.

The choice of acoustic-phonetic features (including scores) will depend on what type of pronunciation errors to detect. Two general measures based on automatic speech recognition (ASR) techniques are log likelihood (LL) and log likelihood ratio (LLR), [3]. A variant of LLR called "Goodness-of-Pronunciation" (GOP) includes models of non-native speech, [2]. We have chosen to use several features; duration, log likelihood and log likelihood ratio, as well as energy. These features are input to a linear classifier trained by linear discriminant analysis (LDA) [4]. We have focused on a general CAPT system where we cannot expect to have enough training data for each L1. Both the ASR models and the LDA-based classifier are trained on native speech.

The outline of the paper is as follows: We present the quantity contrast in Norwegian in section 2 and details on automatic pronunciation evaluation in section 3. The experimental setup is presented in section 4, and is followed by the results in section 5. The discussion and conclusions are given in section 6, and finally directions for further work are given in section 7.

2. Quantity in Norwegian

A special property of the Norwegian phonological system is the quantity contrast where a stressed syllable consists of a short or a long vowel, possibly including consonants in onset and/or coda, [6]. In general the duration is complementary, and phonologically short vowels are followed by a long consonant (VC:) and vice versa (V:C), forming contrasting pairs. A comparison of how this opposition in Norwegian is realized both by natives and non-natives can be found in [7].

Swedish has a similar property, and a study on both perception and production of Swedish prosody including quantity can be found in [8]. Both duration and spectral properties can be important for the perception of VC: versus V:C. A study on how manipulations of duration and formants influence native Swedish perception of the VC:/V:C contrast can be found in [9]. In a CAPT system, both duration (quantity) and spectral (quality) features should therefore be included.

3. Automatic pronunciation evaluation

The L2 pronunciation of an utterance will be assessed by the automatic pronunciation evaluation module. A score is assigned at a suitable level, e.g. phone, dependent on the task.

To evaluate pronunciation at the phone level we need to label and segment the data at the same level. For an operating CAPT system this must be done automatically, e.g. by applying ASR for forced alignment of the waveform and the transcription of the expected pronunciation. Training the LDA-based classifier should be done with the same segmentation procedure to avoid mismatch between training and test.

The utterance must be compared with a representation or model of the "correct" native pronunciation and assign a similarity score. The CAPT exercise will usually include recorded utterances from one or more teacher voices that the user should repeat. These utterances are per definition correct pronunciations and can be used as templates for Dynamic Time Warping type comparison. One problem with using DTW is that the user must imitate the pronunciation and speaking style in order to get a good score. A better and more flexible approach is to use a general model like a Hidden Markov Model (HMM) that includes typical native variation.

After the segmentation the acoustic-phonetic features must be calculated for each segment. For simplicity we have chosen to use the same HMM models both for segmentation and for producing features. However, we are aware of that these two tasks put conflicting demands on the HMM models; accurate boundaries versus consistent acoustic segment content. For acoustic parameterization we have chosen conventional MFCCs including deltas and acceleration.

3.1. LDA-based classifier

The chosen features should produce a final score that reflects the need for feedback on duration. In this work we focus primarily on a binary decision; i.e. long or short. More finely graded decisions are also possible.

As training data are sparse, we need to use a simple classifier in order to be sure to generalize. This is even more important as we use native speech for training and non-native for test. Thus we chose to use a linear classifier. This classifier can be trained by several methods. We chose to use the simple LDA method, i.e. Fisher Discriminant analysis. Thus the score is a real scalar rather than a class index. This enables us to evaluate the chosen features by a score histogram and eventually optimize the threshold or introduce an "uncertainty" class. We may also tune the threshold to weight false rejection as worse than false acceptance.

3.2. Acoustic-phonetic features

In an operating CAPT system we do not know whether the user is able to produce the prompted pronunciation. Corresponding features are therefore computed using both the long vowel HMM and the short vowel HMM. The LDA-based classifier will weight the importance of the different features.

Three main groups of features are considered:
1) Quantity/duration, 2) Quality/spectral content, 3) Energy.

1) Duration: An obvious feature for classifying quantity is the duration of the vowel segment. As the succeeding consonant has a complementary duration this is added as a second duration feature. Segment durations will differ from person to person dependent on the rate-of-speech (ROS). Non-natives are known to have lower ROS than natives. The durations should therefore ideally be normalized. We have instead chosen to keep absolute duration as one feature and include the ratio of the vowel and succeeding consonant durations as a third duration feature. This ratio gives a normalization with ROS and will hopefully boost the contrast between the VC: and V:C syllables.

2) Spectral content: The HMM acoustic model gives a likelihood score of the match between a segment X_k and a HMM phone model Λ_j. This likelihood is a measure of the spectral similarity and is given by

$$LL_k(j) = log[p(X_k|j)] \tag{1}$$

Maximum likelihood training is non-discriminating. The raw log likelihood (LL) from such a system is well suited for comparing scores from different models on the same segment, but not for comparing scores for different segments (from different utterances) against the same model [3]. For CAPT we need to set a fixed threshold for correct/erroneous detection. We have therefore, in addition to LL, used a log likelihood ratio (LLR) measure inspired by confidence scoring.

In confidence scoring the hypothesis that the observation X belongs to class i is deemed to be correct (H_0) or incorrect (H_1) depending on the value of an estimated LLR relative to a threshold. The log likelihood output of the recognizer may be used to model the H_0-hypothesis. So-called "anti-models" are often used as models for the incorrect H_1 classification. Using an average of all other $M-1$ phone models as competitors to form the anti-model, the log likelihood ratio for segment k can be computed using:

$$LLR_k = LL_k(h_0(k)) - \frac{1}{M-1}\sum_{j,j\neq h_0(k)}^{M} LL_k(j) \tag{2}$$

The anti-model can be interpreted as a normalizing factor in order to compare scores for the different segments that we need in a CAPT system. One can alternatively use only the closest competitor, [10], an N-best list, or a limited number of competitors (cohort set). The GOP measure in [2] uses a CAPT-specific cohort set.

3) Energy: Energy is embedded in the spectral score as it is an element of the MFCC vector used as input to the ASR system. The quantity contrast VC: versus V:C can be realized not only in duration or spectral content, but also as energy. We therefore chose to add energy profile as a separate feature.

4. Experimental setup

4.1. Database

Norwegian has 9 pairs of short/long vowels: (/A/,/A:/), (/e/,/e:/), (/i/,/i:/), (/u/,/u:/), (/}/,/}:/), (/y/,/y:/), (/{/,/{:/), (/2/,/2:/), (/O/,/O:/). All phone transcriptions are given in Norwegian SAMPA[1]. Minimal pair two-syllable words were designed where the only difference was the short/long vowel in the first syllable, e.g. the verbs (infinitive form); "lesse" (load) versus "lese" (read). The surrounding consonants will affect the pronunciation and the automatic evaluation system. We therefore also designed a 9-pair set of non-sense words with voiceless plosives as context that is known to be easy to segment: /kVte/ /kV:te/. This gives a total of 36 manuscript words.

For a realistic recording situation an in-house CAPT demo (without feedback) was used as recording set-up. The speakers would see the word spelled out and hear the teacher voice pronounce the word once before they could press a button to record their own voice. The teacher voice had a distinct pronunciation clearly differentiating the vowel length.

[1] http://www.phon.ucl.ac.uk/home/sampa/norweg.htm

The database consists of 13 speakers; 4 Norwegian, 3 Iranian (Farsi speaking), and 6 Chinese (Mandarin speaking). Each speaker recorded the 36 words three times giving a total of 1404 word tokens; 432 native and 972 non-native. Most of the non-natives had a low proficiency level of Norwegian, answering "rarely" when asked how often they spoke Norwegian. Iran and China were chosen as L1 countries because there is an increasing number of immigrants to Norway from these countries and because the languages are very different from each other (and from Norwegian). In [7] these two groups are reported to be among the languages less able to produce the Norwegian quantity contrast.

4.2. Human rating of quantity

Human rating of pronunciation is highly subjective. The task in hand was not to evaluate the non-native pronunciation quality, but rather to find whether native Norwegians would perceive the word as similar to one or the other of the VC:/V:C word pairs.

In these experiments one native Norwegian judge assessed the non-native speech using a 4-class scale enabling the judge to tell whether the decision was sure or unsure. All utterances were rated twice, disagreements are given in Table 1. Most of the conflicts are between "neighboring" classes. Collapsing the score to two classes, VC:/V:C, 31 cases have conflicts giving an agreement rate of 96.9%. The first decision was used giving a total of 418 (43%) perceived VC: and 554 (57%) perceived V:C for the non-native speakers. For the natives we assumed no errors giving 216 VC: and 216 V:C. We did not use any information on the prompted quantity.

	VC:		V:C	
	sure	unsure	unsure	sure
VC: sure	357	34	2	6
VC: unsure	-	8	16	7
V:C unsure	-	-	9	24
V:C sure	-	-	-	509

Table 1: *Disagreements between the two ratings by the judge in perceived VC: or V:C, only non-native pronunciations*

4.3. Training of the ASR system

All speech was sampled using 16 kHz in order to preserve a reasonable bandwidth for the detailed analysis needed in CAPT. We used a fairly standard ASR parameterization of 12 MFCCs as well as the first and second order derivatives. We used normalized energy and cepstral mean subtraction in order to minimize the cross-corpus mismatch. For acceptable segmentation resolution we chose a 5 ms frame shift with a 15 ms window.

The general Norwegian HMM phone models were trained using HTK[2] on about 20 hours of continuous manuscript read speech from about 900 speakers. We used 3-state context-independent acoustic models with 8 Gaussians as a compromise between recognition and segmentation performance.

4.4. Automatic segmentation

The automatic segmentation is crucial for any CAPT system, especially when evaluating quantity. The segmentation should be good enough to detect pronunciation errors, for which consistency is more important than accuracy compared with a human

[2]http://htk.eng.cam.ac.uk/

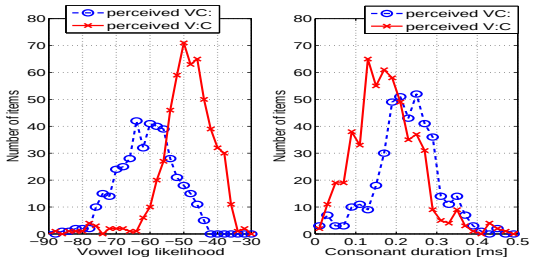

Figure 1: *Histograms using long vowel HMMs on non-native speech for the features vowel log likelihood (left) and succeeding consonant duration (right)*

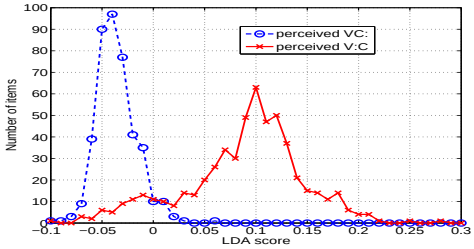

Figure 2: *Histogram of LDA score using all features*

baseline. Nonetheless, we did compare the automatic segmentation on the teacher voice with an available manual one. Of the automatic segmentation boundaries 27% were within 10 ms and 56% within 20 ms tolerance of the manual. Using only long vowel HMMs (wrong for half of the words) gave 25% within 10 ms and 55% within 20 ms. Using only short vowel HMMs (wrong for the other half of the words) gave 24% within 10 ms and 53% within 20 ms. Thus the long vowel HMMs seem to segment short vowels slightly better than the other way round.

4.5. Classifier settings

We have tested 15 acoustic-phonetic features combined with both long and short vowel HMMs, i.e. a total of 30 features:

- 3 duration features: vowel, succeeding consonant, and vowel/consonant duration ratio
- 6 spectral content features: log likelihood of vowel, consonant, and word, log likelihood ratio of vowel, consonant, and word. The LLR is computed using a 5-best list as anti-model.
- 6 energy features: energy profile as 3 values for both the vowel and consonant. The 3 HMM states were used subsegment the two phones into start, mid and end parts for a coarse energy profile.

5. Results

To test the performance on natives we used "leave-one-out" training and testing of the LDA classifier. This gave an average of 5/432 = 1.2% errors and serves as an upper bound on the non-native performance given the features.

The LDA was then trained on all the native speakers and tested on the non-natives. Histograms of two selected features computed for non-native speech are shown in Figure 1. Classification results when using only single feature groups are given

in Table 2. We note that, as expected, vowel duration is more discriminating than the spectral content as measured by LL or LLR. Rather surprisingly, the vowel/consonant duration ratio performs worse than absolute vowel duration. However, this was confirmed by inspecting the average V:/V and C:/C ratios, which turned out to be quite similar for native and non-natives.

Only feature used	Error rate	No. errors
Duration vowel	8.7 %	85
Duration consonant	34.3 %	333
Vowel/consonant duration ratio	15.5 %	151
Log likelihood ratio vowel	28.2 %	274
Log likelihood ratio consonant	47.5 %	462
Log likelihood ratio word	31.4 %	305

Table 2: *Classification errors tested on non-native speech for selected feature pairs (using both short and long vowel HMMs)*

Combining all features gave an error rate of 7.7%, a histogram of the LDA score is shown in Figure 2. Results for the two L1 origins were 10.0% for the Chinese and 3.4% for the Iranians. This reflects the differences in the two L1s compared to L2. To evaluate the correlation between features we removed one feature group at a time. Vowel duration turned out to be the single most important feature, the performance dropped to 9.1% without it. The other durational features gave insignificant contributions when removed one-by-one.

For some features we even observed an increased performance when discarding them, e.g. consonant duration and LL of the word. Our initial guess was that the performance drop was due to a mismatch in the LDA-based classifier caused by the difference in realization of the non-natives compared with the natives. The best combination, given the test set, gave a classification error of 6.7%. To examine this LDA classifier mismatch hypothesis we trained the LDA classifier for each of the two L1s. The "leave-one-out" test results were 9.9% for the Chinese and 4.9% for the Iranians. Thus, this gave no improvement over training the LDA classifier with native speech. One possible explanation is that the non-natives are inconsistent in their pronunciations, giving data not suitable for training.

6. Discussion and Conclusions

We have presented the first results on automatic evaluation of non-native Norwegian pronunciation. The quantity contrast in Norwegian was chosen for the experiments as it is important in order to be understood, but difficult to pronounce for many non-natives. Acoustic-phonetic features representing duration, spectral content and energy were combined using the LDA classifier. A correct classification of 92.3% was achieved.

Absolute vowel duration was clearly the most discriminating feature followed by the "normalized" vowel/consonant duration ratio. The speakers were influenced by the teacher voice, both in pronunciation and perhaps also rate-of-speech. The chosen vocabulary is a realistic type of speech that a CAPT system should be able to score. CAPT is especially efficient for low proficiency level L2-learners where careful word by word pronunciation like this will be one of the exercises. For quantity in continuous speech other features may be more important.

All results are compared with a rating by a single judge. Error analysis revealed that there are different opinions on VC:/VC: perception. There is no objective "ground truth", thus using several judges would give a more robust assessment.

7. Future work

Our experiments have confirmed the expected problem of mismatch in training and test shown by the difference in native/non-native performance. Adaptation on non-native speech is a possible option to reduce this mismatch, [2] and [10]. A CAPT system should be used over a period of time and speaker adaptation can thus be feasible.

Segmentation and segment scoring gives different demands on the ASR systems. Different HMMs tailored for either segmentation or scoring should be investigated further.

The log likelihood ratio did not perform as good as expected. This may be due to either segmentation errors, a too simple anti-model, or that the spectral difference in this case is minor. There are a number of techniques that could be investigated to improve the log likelihood ratio score.

Log likelihood ratio can also be used as a confidence measure to reject utterances. This is useful both if the user says something different than expected and for rejecting pronunciations with other or more errors than the one targeted in the specific exercise.

The final test of a CAPT system is of course whether it helps non-natives to learn Norwegian faster and/or better. A final verification of the usefulness of our quantity module will thus have to wait until a complete CAPT demo is developed.

8. References

[1] B. Granström, "Speech technology for language training and e-inclusion," in *Proc. EuroSpeech-2005*, Lisboa, Portugal, 2005.

[2] S. M. Witt and S. J. Young, "Phone-level pronunciation scoring and assessment for interactive language learning," *Speech Communication*, vol. 30, pp. 95–108, 2000.

[3] Y. Kim, H. Franco, and L. Neumeyer, "Automatic pronunciation scoring of specific phone segments for language instruction," in *Proc. EuroSpeech-1997*, Rhodes, Greece, 1997.

[4] H. Strik, K. Truong, F. de Wet, and C. Cucchiarini, "Comparing classifiers for pronunciation error detection," in *Proc. Interspeech 2007*, Antwerp, Belgium, 2007.

[5] S. Holm, "Intonational and durational contributions to the perception of foreign-accented Norwegian: An experimental phonetic investigation," Ph.D. dissertation, NTNU (Norwegian University of Science and Technology), 2008.

[6] G. Kristoffersen, *The Phonology of Norwegian*. Oxford University Press, 2000.

[7] W. A. van Dommelen, *Non-Native Prosody: Phonetic Description and Teaching Practice*. Mouton de Gruyter, 2007, ch. Temporal patterns in Norwegian as L2, pp. 121–144.

[8] B. Thorén, "The priority of temporal aspects in L2-Swedish prosody: Studies in perception and production," Ph.D. dissertation, Stockholm University, 2008.

[9] D. M. Behne, P. E. Czigler, and K. P. Sullivan, "Swedish quantity and quality: A traditional issue revisited," *Phonum*, vol. 4, pp. 81–83, 1997.

[10] B. Mak, M. Siu, M. Ng, Y.-C. Tam, Y.-C. Chan, K.-W. Chan, K.-Y. Leung, S. Ho, F.-H. Chong, J. Wong, and J. Lo, "PLASER: Pronunciation learning via automatic speech recognition," in *Proc. HLT-NAACL-2003*, Edmonton, Canada, 2003.