NameDat: A database of English proper names spoken by native Norwegians

Line Adde, Torbjørn Svendsen

Department of Electronics and Telecommunications, NTNU Trondheim, NORWAY
line.adde@iet.ntnu.no

Abstract

This paper describes the design and collection of NameDat, a database containing English proper names spoken by native Norwegians. The database was designed to cover the typical acoustic and phonetic variations that appear when Norwegians pronounce English names. The intended use of the database is acoustic and lexical modeling of these phonetic variations. The English names in the database have been enriched with several annotation tiers. The recorded names were selected according to three selection criteria: the familiarity of the name, the expected recognition performance and the coverage of non-native phonemes. The validity of the manual annotations was verified by means of an automatic recognition experiment of non-native names. The experiment showed that the use of the manual transcriptions from NameDat yields an increase in recognition performance over automatically generated transcriptions.

1. Introduction

One of the most difficult and complex problems in Automatic Speech Recognition (ASR) is posed by proper names. There are several elements that can have a detrimental effect on accurate proper name recognition.

Many name recognition applications, such as car navigation or directory assistance applications, contain a considerable number of non-native names. These non-native names are singularly challenging since they have a variety of valid pronunciations. An individual speaker's pronunciation of a non-native name is likely to be influenced by several sociocultural factors. Eklund and Lindström (2001) mention regional background, gender, education and age as important in this regard. Fitt (1995) points out that speakers also tend to use non-native sounds to a varying degree, depending among other things on their knowledge of the name's origin and the origin language. These non-native sounds, then, pose an additional challenge to the ASR engine.

Moreover, predicting reasonable pronunciations for proper names is problematic, considering that these names do not follow conventional pronunciation rules, which makes automatic generation of pronunciation variants a difficult task. Furthermore, manual transcription not only tends to be infeasible from a budgetary perspective, but it also requires expert knowledge in both the native language and languages of other origins present in the corpus.

These issues can be partially remedied by modeling the pronunciation variation either at a lexical level or at an acoustic level. One example of lexical pronunciation variation modeling of proper names is the grapheme-to-phoneme (g2p) phoneme-to-phoneme (p2p) tandem proposed by Yang et al. (2006). Using this scheme, van den Heuvel et al. (2009) were able to automatically generate pronunciation variants for Dutch, English, French and Moroccan proper names which yielded a better performance than the standard variants.

In order to improve the recognition performance of proper names by pronunciation variation modeling, it is crucial to have access to a corpus that contains the different types of variation to be modeled. Several such corpora already exist, for instance the Autonomata Spoken Name Corpus (ASNC) (van den Heuvel et al., 2008) and the cross-lingual database described in (Schaden, 2002). The ASNC contains mostly names of Dutch and Belgian (Flemish) origin, spoken both by native and non-native speakers. Schaden's database consists of European place names spoken by English, French, German and Italian speakers. Both of these contain both speech and transcription data. To date, the only available resource concerning Norwegian pronunciation of non-native names is the Onomastica Consortium (1995) corpus. Unfortunately, this is a purely lexical resource that includes only a single "nativized" transcription for each name in the lexicon, and no recorded speech. These limitations of the Onomastica corpus make it unusable for either lexical or acoustic pronunciation variation modeling. In short, an additional resource for Norwegian is needed.

In this paper we describe the design and collection of NameDat, a small-scale database containing English proper names spoken by native Norwegians. This database was designed as an additional resource to the large vocabulary speech recognition engine SVoG[1]. Its main purpose is to reveal what typical phonetic patterns appear when Norwegians pronounce English names and to improve the recognition performance of English proper names spoken by native Norwegians by modeling the variations seen in the database. The database therefore contains up to seven utterances of every name, spoken by native Norwegian speakers of varying age, gender, education and English proficiency. Each name utterance comes with a detailed annotation made by an experienced phonetician.

In the following sections, the design (section 2), recording (section 3) and annotation (section 4) of the database are described. Finally, section 5 describes two experiments using the NameDat database, and section 6 concludes this paper.

2. Corpus design

The speech data presented in the NameDat corpus was collected from 33 native Norwegian speakers of between 18 and 60 years of age. The speakers were recruited among

[1] http://www.sintef.org/Home/Information-and-Communication-Technology-ICT/Acoustics/Communication-acoustics/SVoG--Large-Vocabulary-Speech-Recognition-for-Norwegian/

colleagues, friends and family. The quality of the recordings are highly dependent on the speakers and their ability and experience with reading aloud. In the corpus, an effort was made to cover some distribution in terms of gender, education and age. As for the parameter of provenance, for such a limited amount of speakers it was unfortunately unfeasible to cover the numerous dialectal regions in Norway. The last design parameter presented in the database is language proficiency, which was determined by means of a self-assessment poll of the speakers. Table 1 gives an overview of the speakers following these parameters.

Criterion	Speakers	
Age	Over 40	Under 40
	12	21
Gender	Male	Female
	17	16
Higher education	Yes	No
	26	7
English proficiency	Intermediate	Good
	10	9
	Very good	Fluent
	11	3

Table 1: Speaker distribution of the NameDat corpus

Each of the 33 speakers read a manuscript consisting of 125 sentences where each sentence contained two names of English origin. There were five different manuscripts in the corpus yielding a total of 1250 unique names. The first four manuscripts were read by seven speakers, while the fifth was read by the remaining five speakers. The manuscripts contained mostly place names from English speaking areas and a smaller part of the corpus contained common US and UK person names.

Three features were especially emphasized in the corpus design. Firstly, it was deemed desirable that the corpus contained both well-known names and names unknown to the speaker. In order to achieve this, two selection criteria were applied, viz. the name's frequency of occurrence in a large text corpus from the news domain, and in case of city names, the city's number of inhabitants. These criteria were taken as a rough indication of the familiarity of the names through the media and travel.

The second feature was to have a considerable amount of "difficult" names in the corpus. We defined a "difficult" name to be a name that a general automatic speech recognizer would have trouble classifying correctly. The Levenshtein distance between an automatically generated transcription and a transcription made by a human expert was used to identify these names. Finally, the third desirable feature was to have a good coverage of non-native sounds in the corpus. Therefore, a special effort was made to include names that feature English phonemes in their pronunciation which are not part of the native Norwegian phoneme alphabet. As such, these particular names supply a good coverage of English sounds that typically have a large pronunciation variation when uttered by Norwegian speakers.

3. Recording

Due to logistic reasons, the recordings were made in two different acoustic environments. The recordings with the majority of the speakers were made in a soundproof acoustic laboratory, while the recordings of the other speakers were made in an office environment.[2] Prior to the recording session the speakers were briefed about the purpose of the project and what was expected of them. They were informed that they would be asked to read 125 Norwegian sentences, all of which contained at least one English name. They were explained that the purpose was not to record the "correct" English pronunciations, but rather to record how they would actually pronounce the names in everyday speech. They were instructed to try to pronounce all names even if they had no idea how to pronounce them.

The recording script was presented to the speaker using the audio recording software Speechrecorder[3]. In order to avoid hesitations, the speakers were instructed to read through the sentence presented on the screen and decide how to pronounce the names in the sentence prior to making a recording.

The recording chain consisted of a Sennheiser HMD 25-1 dynamic headset microphone and Shure FP23 microphone amplifier connected to the line-in port on a MacBook Pro. The signal-to-noise ratio of the recording chain was measured to be 51 dBA and the frequency response of the chain was measured and found to be reasonably flat.

The MacBook Pro was used to digitize the speech. For all recordings a sample rate of 48kHz was used and the samples were stored in 16-bit linear PCM wav format.

4. Broad phonetic annotation

The purpose of the broad phonetic annotation was to document all the different name pronunciations perceived in the recordings and to detect common linguistic features in English proper names spoken by Norwegians. The annotations were later to be used in pronunciation modeling and acoustic modeling of English names, so consistency and accuracy were naturally essential qualities in the annotation process. Currently, two names from 125 carrier sentences have been manually annotated for 19 out of the 33 speakers.

4.1. Annotation format and tools

The phoneme set used for the annotations was in the SAMPA[4] format, with the Norwegian phoneme inventory as the core set. Subsequently, the Norwegian phoneme set was extended with symbols from the British English SAMPA phoneme inventory in order to represent phonemes occurring in English names and loan words which lack an equivalent in the Norwegian inventory. These phonemes are listed in table 2.

The annotations were made in Praat[5] and consisted of five tiers: *auto*, *phone*, *phone comment*, *word*, and *utterance*.

[2] The choice was made out of necessity: 14 speakers were located in the Oslo area, where we had no acoustic laboratory at our disposal.

[3] http://www.phonetik.uni-muenchen.de /Bas/software/speechrecorder/

[4] http://www.phon.ucl.ac.uk/home/sampa

[5] Version 5.0.46, available at http://www.praat.org/

Symbol	Example word
eI	r**ai**se
aU	r**ou**se
@U	n**o**se
r	**wr**ong
w	**w**asp
z	**z**ing
Z	mea**s**ure
D	**th**is
T	**th**in
tS	**ch**in
dZ	**g**in

Table 2: Non-native phoneme extensions

Provisional annotations were available for the whole sentence. For the carrier sentence, the annotations were automatically generated using a Norwegian Text-to-Speech front-end, and for most of the English names, expert transcriptions were available from the Onomastica Consortium (1995) and an in-house pronunciation dictionary. The alignments were obtained using forced alignment.

The provisional annotations were presented to the annotator in the *auto* tier and the corrections were made in the *phone* tier. The annotation was mainly phonetic, but boundaries were corrected where the alignments were clearly misplaced in the provisional annotation. Only the names and name boundaries were corrected. The *phone comment* tier was aligned with the *phone* tier and was used to comment on frequently occurring variations[6].

In the *word* tier, names could be marked as unusable or as mispronunciations. A name was marked as unusable if it contained long pauses or was corrupted by background noise. A name was marked as a mispronunciation if the realization of the name was clearly a reading mistake. For instance, pronouncing the name *'Gilmilnscroft'* as *'Gilmilnsoft'* is obviously a misreading and would be marked as a mispronunciation. However, articulation errors and errors made due to the speaker's insufficient knowledge of English were not marked as mispronunciations. A log file and a pre-defined set of tags were available to the annotator to comment on any uncertainties. The log file can easily be queried by means of the tags.

4.2. Annotation procedure

For consistency reasons, the annotations were performed by one single expert annotator. The annotator was given a set of guidelines and a test session was performed where the annotator received feedback on his annotations. The annotator was instructed to check the provisional transcriptions of the names in the carrier sentences and modify them if necessary. Corrections were made according to general guidelines for Norwegian annotation. In addition, the annotator was instructed to pay special attention to non-native sounds and decide whether or not they were pronounced in a "nativized" manner.

5. Experiments

This section describes two initial experiments using the NameDat database. Section 5.1 describes an automatic name recognition experiment comparing the performance of a system using automatically generated transcriptions with the performance of a system using the manually verified transcriptions in the NameDat database. Two initial multilingual experiments are described in section 5.2.

5.1. Name recognition using automatic and manual transcriptions

Initial isolated word experiments using a dictionary containing 1400 names and the SVoG recognition engine were conducted on a test set of 16 speakers, to evaluate the recognition performance of the names in the database. The SVoG engine employs tri-state, left-right triphone models using 2-64 Gaussian mixture components. Additionally, context-independent monophones with matching topology using 32 Gaussian mixture components were available to the recognition engine. To compensate for acoustic dissimilarities from the recording environment, a global MLLR adaptation was performed on the acoustic models based on the collected material.

Table 3 shows the results of this experiment. The first part of the table lists the results for three baseline systems: with automatically generated Norwegian transcriptions only (*Nor*), with automatically generated English transcriptions only (*En*), and with automatically generated Norwegian and English transcriptions (*NorEn*). The second part of the table gives the result of the system with transcriptions corrected by a phonetician (*Manual*).

Using the manually verified name transcriptions, the error rate decreased 47% relative compared to the highest performing baseline system, which contained both automatically generated Norwegian and English transcriptions. These results confirm that there is indeed a large disagreement between the automatically generated transcriptions and the actual pronunciations. Furthermore, they establish that high quality transcriptions are crucial to obtain satisfactory name recognition performance. Finally, the results indicate that the manual transcriptions are of good quality.

System	Lexicon	WER
Baseline	Nor	52.0%
	En	29.0%
	NorEn	26.0%
Manual	Manual	13.8%

Table 3: Word error rate (WER) for various transcriptions

5.2. Non-native phone recognition

The SVoG recognition engine is a strictly Norwegian resource and is limited to recognizing sounds from the Norwegian phoneme inventory. In view of the considerable amount of non-native phones in the NameDat database, it

[6] Typical comments were e.g. devoicing of voiced phone, uncertain phone identity, phone is realized as an approximant, missing or unknown phone, typical "nativized" pronunciation of an English phone.

was desirable to find a good way of modeling these phones acoustically. In this experiment two different approaches to acoustic modeling of the non-native phonemes in table 2 were investigated. In the first approach, the non-native phonemes were mapped to their acoustically most similar native equivalent. To ascertain acoustic similarity between phonemes, phonetic knowledge of the native language and the non-native language was employed. These models will hereafter be referred to as "nativized" models. In the second approach, the non-native phonemes were trained using the TIMIT[7] database. Both context-dependent and context-independent models were generated. In both cases a 3-state left-right topology with 16 component Gaussian mixture emission densities were used. These models will be referred to as "non-nativized" models.

To evaluate the performance of the proposed approaches, two native phone recognizers were generated: one context-independent (monophones) and one context-dependent (triphones). The monophones and triphones used by these phone recognizers were extracted from the acoustic models used by the SVoG engine. These models were then augmented with nativized and non-nativized monophones and triphones respectively. The resulting nativized and non-nativized models were adapted using a global MLLR transform trained on the collected NameDat corpus to compensate for acoustic dissimilarities in recording environments.

A first multilingual experiment was conducted on a 16-speaker test set using the context-independent phone recognizer augmented with nativized and non-nativized monophones respectively. Table 4 shows that the nativized system performed slightly better than the system augmented with non-nativized models. Investigation of the errors made by the phone recognizer when using the nativized models revealed a wide distribution of classification errors only for the English phone "r". To improve the acoustic representation of this phone, a third set of models was constructed. This model set was identical to the nativized model set apart from the non-native "r" which was trained on the TIMIT database. This yielded a small decrease in phone error rate of 1.1% absolute compared to the nativized system.

Augmented model set	WER
Nativized monophone models	39.2%
Non-nativized monophone models	39.9%
Nativized + "r" monophone models	38.1%

Table 4: Word error rate (WER) of the context-independent phone recognizer augmented with different model sets

A second multilingual experiment was conducted on the same 16-speaker test set using the isolated word recognizer described in the previous section. The models used by the recognizer were augmented with four different sets of models: nativized monophone models, non-nativized monophone models, nativized triphone models and non-nativized triphone models. Table 5 shows the performance of the resulting models. These results reveal that the small performance gap between using the models trained on a non-native database and the models mapped to native models is closed when triphones are used instead of monophones.

Augmented model set	WER
Nativized monophone models	19.5%
Non-nativized monophone models	20.5%
Nativized triphone models	13.8 %
Non-nativized triphone models	13.5%

Table 5: Word error rate (WER) of the context-dependent phone recognizer augmented with different model sets

6. Conclusion and future work

The NameDat database is a small-scale spoken language resource, designed to improve the recognition performance of English proper names spoken by native Norwegians. It is enriched with several annotation tiers providing a detailed annotation for English names. The database has been successfully used in spoken name recognition and in initial multilingual modeling experiments, and is currently being used in pronunciation variation modeling of non-native names.

As shown in section 5, introducing manually verified transcriptions into the lexicon can yield a significant reduction in error rate. These results indicate that modeling the pronunciation variation at a lexical level is likely to be beneficial in terms of error rate reduction. This section further demonstrated that a straightforward substitution of non-native models trained on native speech with non-native models trained on non-native speech does not necessarily result in any significant change in the error rate.

In future work, the database described in this paper will be used to develop more advanced lexical and acoustic pronunciation variation modeling schemes. Further, section 2 reveals that the corpus is somewhat biased towards speakers with a higher education, which may have some impact on the results. The corpus will therefore be extended to include a more balanced set of speakers in terms of education level in the future. Further efforts will also be made to complete the phonetic annotation of the remaining recordings.

The construction of the database described in this paper has been documented in detail, as are the files included in the database and their format. The documented database will be made available for non-commercial purposes through the author or through the forthcoming Norwegian Language bank[8].

7. Acknowledgements

This work has been funded by the Research Council of Norway as part of the research project *Digitale utfordringer: deltakelse og funksjonshemming i informasjonssamfunnet.* The authors would like to thank the employees at Max Manus AS, the colleagues at NTNU, friends and family for their contribution in the making of this corpus. We would also like to thank Dries Tanghe for valuable suggestions and linguistic advice.

[7] http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC93S1

[8] http://www.spraakbanken.uib.no/

8. References

The Onomastica Consortium. 1995. The Onomastica Interlanguage Pronunciation Lexicon. In *Proceedings of the Fourth European Conference on Speech Communication and Technology (EUROSPEECH '95)*, pages 829–832, Madrid, Spain.

Robert Eklund and Anders Lindström. 2001. Xenophones: An investigation of phone set expansion in Swedish and implications for speech recognition and speech synthesis. *Speech Communication*, 35(1-2):81–102.

Susan Fitt. 1995. The Pronunciation Of Unfamiliar Native And Non-Native Town Names. In *Proceedings of the Fourth European Conference on Speech Communication and Technology (EUROSPEECH '95)*, pages 2227–2230, Madrid, Spain.

Stefan Schaden. 2002. A Database for the Analysis of Cross-Lingual Pronunciation Variants of European City Names. In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC'02)*, pages 1277–1283, Las Palmas de Gran Canaria, Spain.

Henk van den Heuvel, Jean-Pierre Martens, Bart D'hoore, Kristof D'hanens, and Nanneke Konings. 2008. The AUTONOMATA Spoken Names Corpus. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco.

Henk van den Heuvel, Bert Réveil, and Jean-Pierre Martens. 2009. Pronunciation-based ASR for names. In *Proceedings of the 10th Annual Conference of the International Speech Communication Association (Interspeech 2009)*, Brighton, UK.

Qian Yang, Jean-Pierre Martens, Nanneke Konings, and Henk van den Heuvel. 2006. Development of a phoneme-to-phoneme (p2p) converter to improve the grapheme-to-phoneme (g2p) conversion of names. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, pages 287–292, Genoa, Italy.