A Longitudinal Bilingual Frisian-Dutch Radio Broadcast Database Designed for Code-Switching Research

Emre Yılmaz[1], Maaike Andringa[2], Sigrid Kingma[2], Jelske Dijkstra[2], Frits van der Kuip[2], Hans Van de Velde[2],Frederik Kampstra[3], Jouke Algra[3], Henk van den Heuvel[1] and David van Leeuwen[1]

[1]Centre for Language and Speech Technology (CLST), Radboud University, Nijmegen, Netherlands
[2]Fryske Akademy, Leeuwarden, Netherlands
[3]Omrop Fryslân, Leeuwarden, Netherlands
{e.yilmaz, h.vandenheuvel, d.vanleeuwen}@let.ru.nl,
{mandringa, skingma, jdijkstra, fvdkuip, hvandevelde}@fryske-akademy.nl,
{frederik.kampstra, jouke.algra}@omropfryslan.nl

Abstract

We present a new speech database containing 18.5 hours of annotated radio broadcasts in the Frisian language. Frisian is mostly spoken in the province Fryslân and it is the second official language of the Netherlands. The recordings are collected from the archives of Omrop Fryslân, the regional public broadcaster of the province Fryslân. The database covers almost a 50-year time span. The native speakers of Frisian are mostly bilingual and often code-switch in daily conversations due to the extensive influence of the Dutch language. Considering the longitudinal and code-switching nature of the data, an appropriate annotation protocol has been designed and the data is manually annotated with the orthographic transcription, speaker identities, dialect information, code-switching details and background noise/music information.

Keywords: Speech database, Frisian, code-switching, longitudinal data, radio broadcasts

1. Introduction

Language contact has been extensively researched in the field of linguistics for more than 60 years (Weinreich, 1953; Auer, 1998; Muysken, 2000; Thomason, 2001). Contact induced language change shows up in the form of phonological, morphological, syntactic and lexical changes consequent to various linguistic phenomena such as word borrowing, interference and relexification. One prominent mechanism that is induced in the interacting languages is code-switching which is defined as the continuous alteration between two languages in a single conversation.

Code-switching is highly noticeable in some minority languages influenced by the majority language or majority languages that have been influenced by globally influential languages such as English and French. Despite the well-established research line in linguistics, robustness of speech-to-text systems against code-switching and other kinds of language switches have recently received some interest (Yu et al., 2004; Lyu et al., 2006; Vu et al., 2012; Wu et al., 2015). Most of this work has been performed on Mandarin-English speech.

In this work, we describe a novel speech database for the West Frisian language. West Frisian is one of the three Frisian languages (together with East and North Frisian spoken in Germany) and it has approximately half a million speakers mostly living in the province Fryslân located in the northwest of the Netherlands. The native speakers of West Frisian (Frisian henceforth) are mostly bilingual and often code-switch in daily conversations due to the extensive influence of the Dutch language (Popkema, 2013).

The Frisian speech data has been collected in the scope of the FAME! (Frisian Audio Mining Enterprise) Project. This project aims to build a spoken document retrieval system for the disclosure of the archives of Omrop Fryslân (Frisian Broadcast) covering a large time span from 1950s to present and a wide variety of topics. Omrop Fryslân is the regional public broadcaster of the province Fryslân. It has a radio station and a TV channel both broadcasting in Frisian and is the main data provider of this project with a radio broadcast archive containing more than 2600 hours of recordings. The Frisian database described in this paper contains a small subset of these radio broadcasts and it is the first spoken Frisian database of high recording and annotation quality. The longitudinal and bilingual nature of the material enables to perform research into language variation in Frisian over years, formal versus informal speech, language change across the life-span, dialectology, code-switching trends, speaker tracking and diarization over a large time period. Moreover, the proposed database will be the main resource for learning the acoustic models that will be incorporated in the spoken document retrieval system.

It is important to note that two kinds of language switches are observed in broadcast data in the absence of segmentation information. Firstly, a speaker may switch language in a conversation (*within-speaker switches*). Secondly, a speaker may be followed by another one speaking in the other language. For instance, the presenter may narrate an interview in Frisian, while several excerpts of a Dutch-speaking interviewee are presented after narration (*between-speaker switches*). The former type is in line with the definition of CS phenomenon in linguistics, while the latter occurs due to the broadcast nature of the data. Both type of switches pose a challenge to the ASR systems and have to be handled carefully during recognition.

The rest of the paper is organized as follows. The next section lists some prior work describing the efforts to collect longitudinal and code-switching speech data. Then, we explain the annotation process and the protocol by presenting

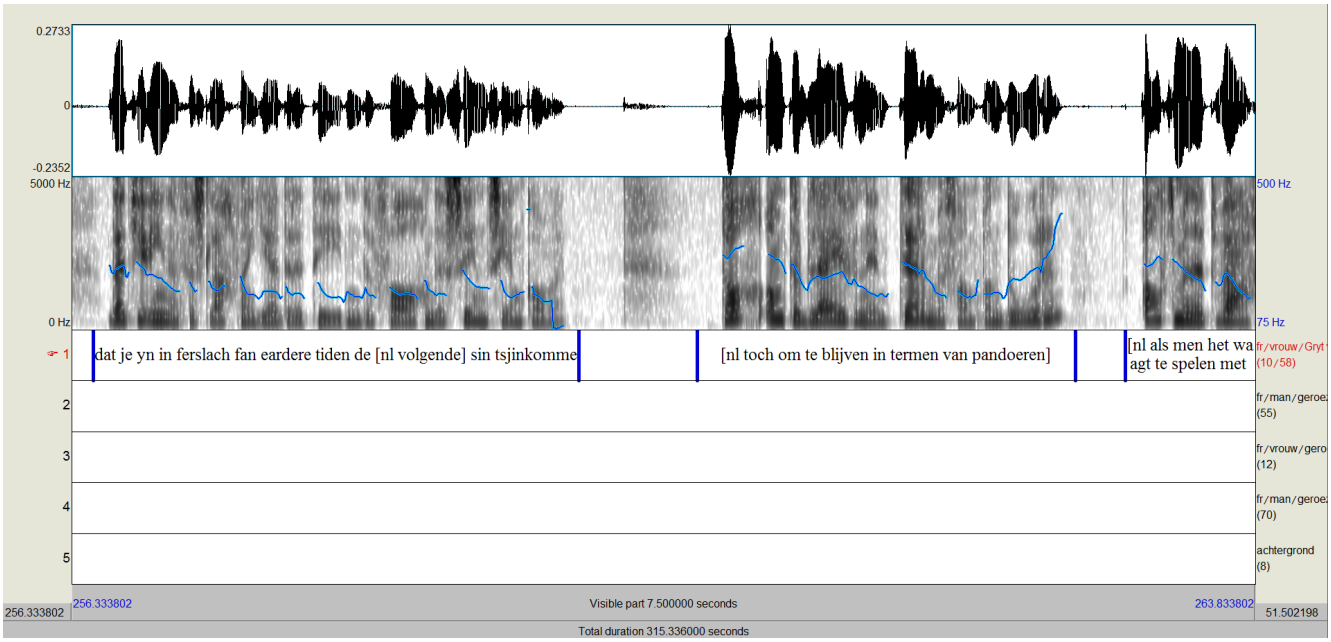

Figure 1: A textgrid file containing a speech segment with code-switching

some examples from the database. Finally, we will provide some statistics from the database detailing the code-switching and speaker information before the concluding the paper.

2. Related Work

Investigation of code-switching in the context of automatic speech recognition research has become viable with several code-switching databases that have been proposed in the last years (Lyu et al., 2015; Li and Fung, 2012; Dey and Fung, 2014; Chan et al., 2005; Imseng et al., 2012). These databases contain recordings of Mandarin-English, Hindi-English, Cantonese-English and French-German code-switching speech data. The automatic speech recognition systems applied on these data use bilingual pronunciation dictionary and language models to be able to cope with the language switch. Moreover, several language identification techniques are adopted to label the speech segments with the spoken language and perform accurate acoustic and language modeling based on these labels.

Longitudinal speech databases have been designed for exploring the effects of speaker ageing on the accuracy of automatic speaker recognition systems and accent development of non-native speakers of English (Brandshein et al., 2010; Orr et al., 2011; Kelly et al., 2013). The recently organized MGB Challenge[1] also focuses on longitudinal speech-to-text transcription and speaker diarization in the episodes of the same series of programs from the British Broadcasting Corporation (BBC). The challenge participants have been asked to identify the common speakers that appear multiple times over time in the different episodes of the same program.

Language identification and/or diarization is a relevant task for the automatic speech recognition (ASR) of code-switching speech (Weiner et al., 2012; Lyu et al., 2013; Yeong and Tan, 2014; Mabokela et al., 2014). One fundamental approach is to label speech frames with the spoken language and perform recognition of each language separately using a monolingual ASR system at the back-end. These systems have the tendency to suffer from error-propagation between the language identification front-end and ASR back-end, since language identification is still a challenging problem especially in case of intra-sentence CS. To alleviate this problem, single-pass ASR approaches, which do not directly incorporate a language identification system, have also been proposed yielding promising results (Lyu et al., 2006; Lyudovyk and Pylypenko, 2014).

3. Annotation Details

The radio broadcast recordings have been manually annotated by two native Frisian speakers. The annotation protocol designed for this code-switching Frisian data includes three kinds of information: the orthographic transcription containing the uttered words, speaker details such as the gender, dialect, name (if known) and spoken language information. To be able to assess the impact of background noise and/or music on the recognition accuracy, the segments containing background noise/music are also labeled. In order to get more precise information about the speaker details, the meta-information of all available radio broadcasts is also provided together with the recordings.

The annotation has been performed using the PRAAT software (Boersma and Weenink, 2015) and the annotated information is stored in textgrid files. An example textgrid file containing code-switching speech is shown in Figure 1. The speaker and spoken language information is stored in the tier names and the orthographic transcription and language switching information are stored in the tiers. The tier names are structured to contain all available informa-

[1] http://www.mgb-challenge.org/

tion about the speaker and spoken language in the format given below.

```
language-dialect/gender/speaker name
```

Focusing on the challenges introduced by the language interaction between the Frisian and Dutch language for speech-to-text systems, the annotation protocol does not distinguish between different types of language interaction. The switches in the spoken language are marked in the parenthesis including the acronym of the language. For clarity, we demonstrate how the language switches are marked with an example sentence extracted from the database. The Frisian speaker uttering the sentence below switches twice to Dutch and the Dutch loan words are marked with `[nl ...]`.

```
wy prate [nl namelijk] mei Marijke
Nicolai en it is folle [nl ernstiger]
```
[2]

Similarly, when a Dutch speaker switches to Frisian, the Frisian words/sentences are marked using `[fr ...]`. Finally, we use `[fr-nl ...]` for marking the words that can neither be classified as Dutch nor as Frisian. These kind of words include Dutch words pronounced according to Frisian pronunciation rules, Dutch verbs conjugated according to Frisian grammar, compound words consisting of a Frisian and a Dutch word.

The annotation procedure is organized in multiple stages to enhance the annotation quality. In the first stage, the annotators and the technical staff were in close contact for customizing the annotation protocol to meet the requirements of target research questions. The transcription time in this stage was as low as 27.8 hours for one hour of annotated speech. In the later stages, the transcription time gradually reduced to 16.7 hours for one hour of annotated speech. Every annotated audio segment is cross-checked by the other annotator to avoid systematic annotation errors and to verify the quality of the annotation.

4. Data

The total duration of the manually annotated radio broadcasts sums up to 18 hours, 33 minutes and 57 seconds. The stereo audio data has a sampling frequency of 48 kHz and 16-bit resolution per sample. The database consists of 203 audio segments of approximately 5 minutes long extracted from various radio programs recorded in different years. The content of the recordings are very diverse including radio programs about culture, history, literature, sports, nature, agriculture, politics, society and languages.

The total amount of audio segments containing speech is approximately equal to 14 hours. This data is divided into training, development and test sets to be able to perform ASR experiments. The training data of the database comprises of 8.5 hours and 3 hours of speech from Frisian and Dutch speakers respectively. The development and test sets each consist of 1 hour of speech from Frisian speakers and 20 minutes of speech from Dutch speakers.

[2]English translation: "We talk indeed with Marijke Nicolai and it is far more serious."

Speaker name	Gender	# of Appear.
E. Ennema	M	24
G. van Duinen	F	23
S. Tigchelaar	F	18
K. Bies	F	14
E. Lok	M	13
G. de Vries	F	10
G. de Vries	M	9
B. de Groot	M	9
R. Koster	M	8
A. van der Mark	F	8
A. de Hoop	F	7
R. Tolsma	M	6
S. Dijkstra	M	5
M. van Kammen	M	4
H. te Biesebeek	M	4
H. Bakker	M	4
S. van der Veen	F	3
L. Dykstra	F	3
K. Wielinga	M	3
K. Gildemacher	M	3
J. van der Zee	M	3

Table 1: List of speakers appearing multiple times in the proposed database

Thanks to the efforts of our data provider, it is ensured that the database contains several speakers such as program presenters and celebrities appearing multiple times in different recordings. The available meta-information helped the annotators to identify these speakers and mark them either using their names or the same label (if the name is not known). There are 309 identified speakers in the FAME! speech database, 21 of whom appear at least 3 times in the database. The list of these speakers and their appearance counts are given in Table 1. These speakers are mostly program presenters and celebrities appearing multiple times in different recordings over years. There are 233 unidentified speakers due to lack of meta-information.

The total number of word- and sentence-level code-switching cases in the FAME! speech database is equal to 3837. These switches are mostly performed by the Frisian speakers as they often use Dutch words or sentences while speaking in Frisian. These cases comprise about 75.6% of the all switches. The opposite case, i.e., a Dutch speaker using Frisian words or sentences, occurs much less accounting for 2.5% of all switches. This is expected as it is not common practice for Dutch speakers to switch between Dutch and Frisian. In the rest of the cases, the speakers use a *mixed-word* which is neither Frisian nor Dutch. The training, development and test sets contain 2756, 671 and 410 language switching cases respectively.

5. Conclusions

We have detailed our efforts towards building a Frisian speech database containing 18.5 hours of radio broadcasts collected from the archives of the local broadcaster. The recordings have been manually annotated according to a

dedicated annotation protocol which is designed considering the code-switching nature of the Frisian language. The database will be made publicly available in the near future and is expected to contribute to several research fields such as code-switching analysis in speech-to-text systems, automatic language identification, speaker ageing effects in speaker recognition and diarization.

6. Acknowledgments

This research is funded by the NWO research grant with Ref. no. 314-99-119 (Frisian Audio Mining Enterprise).

7. Bibliographical References

Auer, P. (1998). *Code-switching in Conversation: Language, Interaction and Identity*. London, Routledge.

Boersma, P. and Weenink, D., (2015). *Praat: doing phonetics by computer, Version 5.4.12*.

Brandshein, L., Graf, D., Cieri, C., Walker, K., Caruso, C., and Neely, A. (2010). Graybeard-voice and aging. In *Proc. of the International Conference on Language Resources and Evaluation (LREC)*, pages 2437–2440.

Chan, J. Y. C., Ching, P. C., and Lee, T. (2005). Development of a Cantonese-English code-mixing speech corpus. In *Proc. European Conference on Speech Communication and Technology*, pages 1533–1536.

Dey, A. and Fung, P. (2014). A Hindi-English code-switching corpus. In *Proc. of the International Conference on Language Resources and Evaluation (LREC)*, pages 2410–2413.

Imseng, D., Bourlard, H., Caesar, H., Garner, P. N., Lecorv, G., and Nanchen, A. (2012). Mediaparl: Bilingual mixed language accented speech database. In *IEEE Spoken Language Technology Workshop (SLT)*, pages 263–268, Dec.

Kelly, F., Brümmer, N., and Harte, N. (2013). Eigenageing compensation for speaker verification. In *Proc. INTERSPEECH*, pages 1624–1628.

Li, Y. and Fung, P. (2012). Code switching language model with translation constraint for mixed language speech recognition. In *Proc. COLING*, pages 1671–1680, Dec.

Lyu, D.-C., Lyu, R.-Y., Chiang, Y.-C., and Hsu, C.-N. (2006). Speech recognition on code-switching among the Chinese dialects. In *Proc. ICASSP*, volume 1, pages 1105–1108, May.

Lyu, D.-C., Chng, E.-S., and Li, H. (2013). Language diarization for code-switch conversational speech. In *Proc. ICASSP*, pages 7314–7318, May.

Lyu, D.-C., Tan, T.-P., Chng, E.-S., and Li, H. (2015). Mandarin-English code-switching speech corpus in South-East Asia: SEAME. *Lang. Resour. Eval.*, 49(3):581–600, Sep.

Lyudovyk, T. and Pylypenko, V. (2014). Code-switching speech recognition for closely related languages. In *Proc. SLTU*, pages 188–193, May.

Mabokela, K. R., Manamela, M. J., and Manaileng, M. (2014). Modeling code-switching speech on under-resourced languages for language identification. In *Proc. SLTU*, pages 225–230.

Muysken, P. C. (2000). *Bilingual Speech: A Typology of Code-mixing*. Cambridge University Press.

Orr, R., Quen, H., van Beek, R., Diefenbach, T., van Leeuwen, D. A., and Huijbregts, M. (2011). An international English speech corpus for longitudinal study of accent development. In *Proc. INTERSPEECH*, pages 1889–1892.

Popkema, J. (2013). *, Frisian Grammar: The Basics*. Afûk, Leeuwarden.

Thomason, S. (2001). *Language Contact - An Introduction*. Edinburgh University Press.

Vu, N. T., Lyu, D.-C., Weiner, J., Telaar, D., Schlippe, T., Blaicher, F., Chng, E.-S., Schultz, T., and Li, H. (2012). A first speech recognition system for Mandarin-English code-switch conversational speech. In *Proc. ICASSP*, pages 4889–4892, March.

Weiner, J., Vu, N. T., Telaar, D., Metze, F., Schultz, T., Lyu, D.-C., Chng, E.-S., and Li, H. (2012). Integration of language identification into a recognition system for spoken conversations containing code-switches. In *Proc. SLTU*, May.

Weinreich, U. (1953). *Languages in Contact : Findings and Problems*. New York, Linguistic Circle.

Wu, C.-H., Shen, H.-P., and Hsu, C.-S. (2015). Code-switching event detection by using a latent language space model and the delta-bayesian information criterion. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 23(11):1892–1903, Nov.

Yeong, Y.-L. and Tan, T.-P. (2014). Language identification of code switching sentences and multilingual sentences of under-resourced languages by using multi structural word information. In *Proc. INTERSPEECH*, pages 3052–3055, Sept.

Yu, S., Zhang, S., and Xu, B. (2004). Chinese-English bilingual phone modeling for cross-language speech recognition. In *Proc. ICASSP*, volume 1, pages 917–920, May.