# Translation and Prosody in Swiss Languages

Philip N. Garner, Rob Clark, Jean-Philippe Goldman, Pierre-Edouard Honnet, Maria Ivanova, Alexandros Lazaridis, Hui Liang, Beat Pfister, Manuel Sam Ribeiro, Eric Wehrli, Junichi Yamagishi

Idiap Research Institute, University of Geneva,
ETH Zurich, University of Edinburgh
<Phil.Garner@idiap.ch>

**Résumé**

*Le projet SIWIS (Spoken Interaction with Interpretation in Switzerland) a pour but principal la traduction parole-à-parole. L'originalité du projet tient dans la transmission des propriétés voco-prosodiques du locuteur original. Autrement dit, la traduction ne véhicule pas seulement le contenu linguistique, mais aussi des traits d'émotion et d'intention linguistique propres au locuteur. Nous décrivons ici le cadre du projet, les approches de recherche et les premiers résultats, dont la conception et constitution d'un corpus oral bilingue qui vise à l'entrainement et l'évaluation de systèmes de reconnaissance et de synthèse de la parole, avec une composante d'accent regional suisse.*

***Keywords:*** *automatic speech recognition, text-to-speech synthesis, speech-to-speech translation, prosody*

## 1. Introduction

Recently, global interest in speech-to-speech translation (S2ST) has increased. This is partially due to the recognition of the "language barrier" by the EU, and also partially due to the ready availability of the component technologies, often as APIs. For instance, many (larger) organisations have recently demonstrated S2ST in the context of smartphones. A recent EU project, EMIME[1] (Wester et al., 2010), aimed to close the gap between the technologies responsible for automatic speech recognition (ASR) and text-to-speech (TTS) synthesis, and that remains an expected impact of the recent EU Horizon 2020 call.

EMIME capitalised on the (relatively) recent convergence of ASR and TTS technology afforded by *statistical parametric* TTS. Rather than rely on overlap-add (OLA) techniques, statistical TTS (Zen, Tokuda, & Black, 2009) is based on the hidden Markov model (HMM) technology that has been used in ASR since the mid-1980s. The convergence allowed techniques developed for ASR to be used almost unchanged in TTS. Notably, the linear transform based adaptation of Leggetter

[1] http://www.emime.org/

and Woodland (1994) could be used to create a great variety of TTS voices using comparatively small amounts of training data (Yamagishi et al., 2009, 2010). The contribution of EMIME, building on work of Wu, Nankaku, and Tokuda (2009), was to allow this adaptation to happen cross-lingually (Liang, Dines, & Saheer, 2010; Liang & Dines, 2010). This resulted in a S2ST system where the output voice in L2 had the characteristics of the input voice in L1.

SIWIS is a Swiss national science foundation funded project in S2ST. It brings together expertise in the component technologies of S2ST, being ASR, translation and TTS, along with representative language groups of Switzerland (including, pragmatically, English). SIWIS continues in the vein of EMIME, but with a particular focus on prosody. Prosody is an important aspect of TTS, but often ignored in both ASR and translation. One goal is to extend current ASR technology such that it can extract prosody; we also intend to pass prosodic events through the translation process such that they can be reflected in the resulting TTS.

Switzerland brings the availability of bilingual speakers. Such speakers allow recordings to be made in two languages, leading to two key capabilities:

– Statistical models can be constructed that can isolate *speaker* characteristics from *language* characteristics. This is an issue that was identified in EMIME, but could not be addressed owing to lack of suitable data.

– Synthesis in L2 based on data in L1 can be more easily evaluated because the ground truth L2 is available from the target speaker. Evaluators can focus on speaker similarity evaluation, without being encumbered by language differences.

The Swiss locality itself brings a third capability, that of an abundance of native speakers capable of evaluating monolingual TTS.

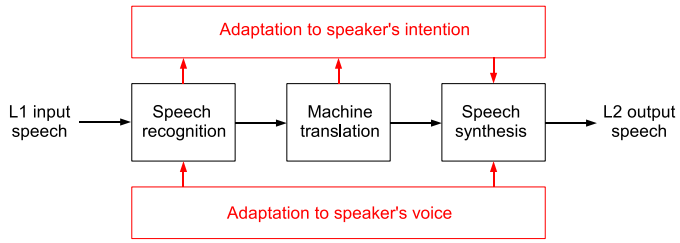


Figure 1: *The black parts are a traditional S2ST system. EMIME added the lower path in grey; SIWIS will focus on the upper path*

Although one goal of SIWIS is to progress speaker adaptation technology, the main goal is to add the extra dimension of *intent*. The black

parts of figure 1 illustrate a traditional S2ST system. Notice that no voice characteristics are transferred. The grey part underneath was the focus of EMIME, where the goal was to use adaptation techniques from ASR to adapt the synthesis in L2.

The goal of SIWIS is concerned with the upper grey part of figure 1: transfer of intent. Intent is not carried in the speaker characteristics; rather, it is a function of emotion, and hence prosody. This amounts to three tasks:

1. Extraction of prosody at the ASR stage.
2. Translation of prosody along with the usual words.
3. Synthesis of the translation with prosody.

The state of the art in this chain is somewhat sparse. Certainly there have been other large projects addressing S2ST such as VERBMOBIL[2], TC-STAR[3] and Gale[4]. Pipeline systems have coupled the component technologies (Ney, 1999; Gao, 2003). Noth, Batliner, Kiessling, Kompe, and Niemann (2000) describe how prosody can be used to speed syntactic parsing. However, we are unaware of systems that handle prosody end to end.

In the following sections, we give an overview of how SIWIS is addressing the above goals.

## 2. SIWIS bilingual database

The SIWIS bilingual database gathers 84 bilingual speakers, reading aloud about 180 prompts in two languages among the four of the project (i.e. French, English, German, and Italian). Each of the 6 possible language pairs is represented by 14 speakers (7 males and 7 females). This corpus has several purposes: training data for ARS and TTS, cross-language speaker adaptation, evaluation data for TTS, ad hoc data for prosody studies, in particular focus.

The recruitment was done mainly in Geneva through the University and the international organizations. Advertisement with wall notices and mailing-list pointed people to a web page on which they could record themselves in 2, 3 or 4 languages. The task was to read a short passage of the book "Le Petit Prince" in every language they applied for. Each recording was judged by 3 native speakers on a scale of 0 to 3 (0=clearly accented, 1=noticeable accent, 2=very slight accent, 3=no accent at all). The speakers with a mean score above 2.5 were selected (e.g all "3"s, or possibly one "2") representing about 40% of the candidates.

[2] http://verbmobil.dfki.de/overview-us.html

[3] http://www.tc-star.org

[4] http://www.darpa.mil/ipto/programs/gale/gale.asp

A minority of them were considered as trilingual. None of them were selected as quadrilingual.

The selected speakers were recorded in a booth at the University of Geneva for two series of about 180 prompts (three series for the trilinguals) and paid 60.- (90.- for trilinguals). They were shown each prompt on the screen, could take time to read the prompt silently and triggered themselves the recording. They repeated prompts for which the reading was hesitant. Each series is divided in 5 parts:

– 25 from Europarl (among which 5 questions). The meaning is the same across the 4 languages;

– 25 from Europarl (same prompts as above) with one same focused word across languages;

– 100 from newpapers (80 declaratives, 20 questions);

– 20 semantically unpredictable sentences;

– 10 sentences from the book "The little prince / Le petit prince" presented as a paragraph.

Each language has actually three different sets of prompts, excepted for the fifth part which is the same for the three sets. All in all, the SIWIS bilingual database includes more than 15000 prompts from 84 speakers (actually from 44 different individual speakers among which 20 are trilinguals). As of June 2014, the corpus is completed at 50%.

**3. Interface between ASR and Machine Translation**

Given that the text to be translated comes from ASR, the translation engine is being extended to deal with alternatives presented by the recogniser. A syntactic parser tries to help to select the right alternatives in order to ensure a better translation. A significant amount of research work that focuses on improving ASR output by post-processing it already exists, and some successful results have been reported (Huet, Gravier, & Sébillot, 2010; Bassil & Alwani, 2012; Bassil & Semaan, 2012; Feld, Momtazi, Freigang, Klakow, & Müller, 2012; Huet, Gravier, & Sébillot, 2008).

The ASR results based on our development data suggested that some errors occuring in the ASR output might be avoided by syntactic analysis. Some occurrences of ungrammatical sequences of words were observed. To account for these cases, we explored the lower-ranked hypotheses produced by the ASR system.

We try to improve the ASR output by reordering the N-best list produced by the ASR system via syntactic filtering. The proposed interface is being developed between an ASR system and a rule-based machine translation (MT) system (Wehrli, Nerima, & Scherrer, 2009). An important component of the MT system is its parser (Wehrli, 2007). The

hypotheses of each utterance are submitted to the parser in descending order, starting with the 1-best ASR hypothesis. The parser goes through the hypotheses one by one until it can parse a hypothesis. Then it continues further with the next utterance. If no hypothesis can be parsed, the default 1-best ASR hypothesis is selected.

The output of an ASR recognition system lacks certain features from a standard written document such as punctuation and capitalization. Therefore we needed to adapt our parser to this output. Many syntactic rules use punctuation (e.g., commas, dashes, parentheses, etc.) to deal with the structure of more complex sentences (e.g., containing, for example, relative clauses). Due to all these issues, we needed to relax the parser, such that some syntactic rules could be used in a less strict way.

Our results up to now are based on English using the Wall Street Journal data (Paul & Baker, 1992). We report similar results on word recognition, but decrease in sentence recognition.

The results obtained with the default 1-best ASR hypothesis, as well as with the reordering of the hypotheses via syntactic filtering, are shown in Table 1. The column with label *U. % Corr* shows the percentage of complete utterances, which were recognized correctly. The other two columns *W. % Corr* and *W. Acc* show the word recognition and accuracy statistics for the individual words (Young, Odell, Ollason, Valtchev, & Woodland, 1997). The word recognition results obtained from the reordering of the hypotheses are similar to the 1-best ASR results. The complete utterances recognition, however, decreases by 4.5%.

| | U. %Corr | W. %Corr | W. Acc |
|---|---|---|---|
| 1-b. ASR | 42.00 | 93.41 | 91.66 |
| 10-best | 37.50 | 93.11 | 91.14 |
| 20-best | 37.00 | 92.89 | 90.95 |

Table 1: *Comparison of ASR 1-best and hypotheses reordering*

The decreased score on complete utterance recognition results from the fact that the parser skips the correct hypothesis, because it cannot parse it, and chooses another lower-ranked hypothesis. We hope that with a better parser performance these errors could be avoided.

**4. Swiss French accents**

It is commonly known that there is no big difference between French spoken in France and French spoken in Switzerland. In particular, the pronunciation of standard French, as defined by Morin (2000), and Swiss accent are close and present only few differences at segmental level (Métral, 1977). However, at the prosodic level, several differences are highlighted in the literature and Sertling Miller (2007) and Schwab

et al. (2012); Schwab and Racine (2013) give an overview of some of these differences.

Furthermore, among the Swiss French regional accents, the variations in speech occur in both segmental and suprasegmental domains. These differences are subtle and thus can not be considered as phonological differences. The variations are mainly focused on the speaking style, i.e. different rhythm and pitch variations, rather than on the pronunciation of the words (Lodge, 1993; Racine, Schwab, & Detey, 2013; Woehrling & de Mareüil, 2006), making the task of regional accent identification or synthesis even more difficult.

Preliminary work on automatically recognizing the speaker's accent among regional Swiss French accents from four different regions of Switzerland, was done recently by Lazaridis et al. (2014). An attempt was made to cast this task as a biometric identification problem, relying on techniques which were first introduced in speaker recognition and then successfully applied for several audio processing problems. To achieve this goal, a generative probabilistic framework for classification based on *Gaussian mixture modelling* (GMM), was implemented. Two GMM-based algorithms were investigated: (1) the baseline technique of universal background modelling (UBM) followed by maximum-a-posteriori (MAP) adaptation (Reynolds, Quatieri, & Dunn, 2000) and (2) the total variability (i-vectors) modelling (Dehak, Kenny, Dehak, Dumouchel, & Ouellet, 2011).

| System | GE | MA | NE | NY | Total Accuracy |
|---|---|---|---|---|---|
| GMM | 23.7% | 38.5% | 19.6% | 54.6% | 33.4% |
| TV-SVM | 35.1% | 32.9% | 25.7% | 63.4% | 38.5% |

Table 2: *Performance summary: This table reports the accuracy of the GMM and TV-SVM systems on Swiss French regional accent identification*

In Table 2, the accuracy of the two systems is shown along with the total accuracy of each system. As can be seen, the TV-SVM system outperforms the GMM-based system in the 3 out of 4 accents. A relative improvement of 15.3% in the overall accent identification accuracy was achieved by TV-SVM over the GMM-based system.

In our attempt to synthesize regional accents from Switzerland, a preliminary study was conducted on prosody in Swiss French accent perception by Honnet, Lazaridis, Goldman, and Garner (2014).

The main idea was to investigate how prosody helps in perceiving Swiss accent. For that, we used standard French TTS models and Swiss accented French data to combine standard French spectral parameters (as for pronunciation) and Swiss prosody. In particular we used duration and intonation from Swiss French data to observe their effect on accent perception. Figure 2 summarizes the experiment.

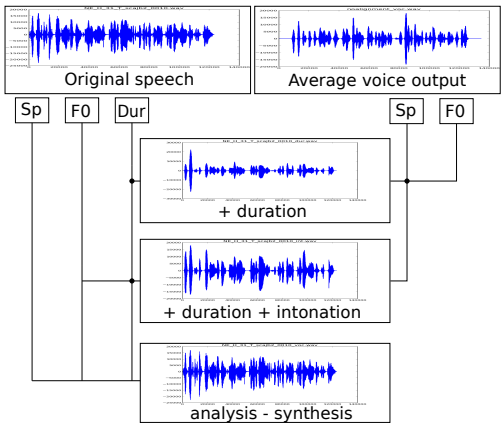


Figure 2: *Experimental setup. Features are Sp for spectrum, F0 for fundamental frequency, Dur for duration*

The degree of accent (using French from France as a reference) was then assessed by native French speakers through a listening test. The listeners had to rate the degree of accent of three files per speaker (twelve speakers: ten Swiss from five different cities and two Parisians): one file where the duration information is given, one where duration and intonation are given, and the original file.

The results obtained showed that, starting from an average standard French male synthetic voice and adding duration information extracted from Swiss French male speakers resulted in a degree of accent closer to the original speech. Adding duration and intonation resulted in a degree even closer to the original speech. However, for some cases (mainly strongly accented speakers), adding the prosodic features resulted in an accent still much lower than the original speech.

## 5. Multi-Level Modelling of Prosody

It is widely agreed that prosody is inherently supra-segmental. In segmental phonology, modifying the identity of one segment ('pin'/'bin'), or the position of the lexical stress ('inTERN' (verb) / 'INtern' (noun)) changes the meaning of the lexical item and its overall context. However, if we change the sequence of segments in an utterance, listeners are still capable of discerning the same melody and rhythm. Prosodic variations are often seen as a layer that lies on top of a sequence of segments and their properties.

However, current speech synthesizers (Zen et al., 2007, 2009) still model prosody using short-term methodologies, which have been inspired and inherited by segmental modeling. The long-term dependencies are captured somewhat implicitly through the use of rich context

models, defined by a set of shallow supra-segmental features (number of syllables and words in prosodic phrase and utterance, distance to next and previous stress and pitch accent, etc). But even though the features are present, modelling still takes place at frame level.

We therefore intend to explore the supra-segmental nature of prosodic variations by investigating multi-level modelling approaches. For example, model short-term pitch and duration variations at lower levels (such as frame, phone, or syllable), and long-term pitch and duration variations at higher levels (such as word, prosodic phrase, or utterance). Recent approaches have shown that superpositional (Zen & Braunschweiler, 2009; Stan & Giurgiu, 2011) or joint (Latorre & Akamine, 2008; Qian, Wu, Gao, & Soong, 2011) multi-level models are promising.

Initial experiments regarding f0 indicate that modelling high frequencies separately from low frequencies using a wavelet transform similar to that of Suni et al. (2013) yields encouraging results.

**6. Syntax Tree-Based Prosody Modelling**

A prosody generation module for concatenative TTS that makes direct use of syntax trees to predict duration and pitch had been developed (Hoffmann & Pfister, 2012). In brief, two stages are involved in this module: (i) a prosody contour at the word level is generated from the syntax tree of a sentence; (ii) the prosody contour, together with contextual information about phones in the sentence, is further processed by ANN to yield concrete prosodic features ($F_0$ and duration) for this sentence. In effect, the word-level prosody contour mentioned in (i) is just a description of concrete prosodic features, and is portrayed as a sequence of vectors. Each of these word-level vectors consists of seven elements[5] and corresponds to a single word in the sentence.

Since the syntax tree-based prosody generation module was previously developed only on speech data from a single speaker, we are much interested in examining the possibility of this module capturing speaker-independent prosodic information. The difference between the two stages explained above implies stage (i) could be primarily handling the general tendency of prosody variation of a sentence and that stage (ii) could be largely handling more specific, local patterns of prosody. Consequently, we intend to find out the extent of speaker-independence of word-level prosody contours by checking how applicable they are to ANNs trained on speech data from a different speaker (Liang, Hoffmann, & Pfister, 2014).

We experimented on the German language. Speech samples pre-

[5] Normalised $F_0$ mean, degree of flatness of the $F_0$ curve of the word, gradient of the $F_0$ curve of the word, length of the pause before/after the word, normalised duration of the word, gradient of duration of the word.

sented to listeners[6] were generated by an HMM-based speech synthesiser whose conventional prosody prediction mechanism was replaced with our syntax tree-based prosody generation module. ANNs were always trained in the speaker-specific manner, while word-level prosody contours were trained on speech data in various sepakers' voices. Subjective evaluations of cross-speaker combinations of the ANNs and the word-level prosody contours suggested the prosody contours mainly captured speaker-independent prosodic information and could work well with the ANNs trained on speaker-specific data in a different voice.

**Acknowledgements**

This work has been conducted with the support of the Swiss NSF under grant CRSII2 141903: Spoken Interaction with Interpretation in Switzerland (SIWIS).

**References**

Bassil, Y., & Alwani, M. (2012). Post-editing error correction algorithm for speech recognition using bing spelling suggestion. *International Journal of Advanced Computer Science and Applications*, *3*, 95–101.

Bassil, Y., & Semaan, P. (2012). ASR context-sensitive error correction based on Microsoft N-Gram dataset. *Journal of computing*, *4*, 34–42.

Dehak, N., Kenny, P., Dehak, R., Dumouchel, P., & Ouellet, P. (2011). Front-end factor analysis for speaker verification. *IEEE Transactions on Audio, Speech and Language Processing*, *19*(4), 788–798.

Feld, M., Momtazi, S., Freigang, F., Klakow, D., & Müller, C. A. (2012). Mobile texting: can post-ASR correction solve the issues? An experimental study on gain vs. costs. In C. Duarte, L. Carriço, J. A. Jorge, S. L. Oviatt, & D. Gonçalves (Eds), *IUI'2012* (pp. 37–40). Lisbon: ACM.

Gao, Y. (2003). Coupling vs. unifying: Modeling techniques for speech-to-speech translation. In *Eurospeech* (pp. 365–368). Geneva, Switzerland.

Hoffmann, S., & Pfister, B. (2012). Employing sentence structure: Syntax trees as prosody generators. In *Proceedings of Interspeech* (pp. 470–473).

Honnet, P.-E., Lazaridis, A., Goldman, J.-P., & Garner, P. N. (2014). Prosody in Swiss French accents: Investigation using analysis by synthesis. In *Proceedings of Speech Prosody.* Dublin, Ireland.

Huet, S., Gravier, G., & Sébillot, P. (2008). Morphosyntactic resources for automatic speech recognition. In *Proceedings of the 6th international conference on language resources and evaluation (LREC'08).* European Language Resources Association (ELRA).

Huet, S., Gravier, G., & Sébillot, P. (2010). Morpho-syntactic post-processing of N-best lists for improved French automatic speech recognition. *Computer Speech & Language*, *24*(4), 663–684.

[6] All are native German speakers.

Latorre, J., & Akamine, M. (2008). Multilevel parametric-base f0 model for speech synthesis. In *Proceedings of Interspeech* (pp. 2274–2277).

Lazaridis, A., Khoury, E., Goldman, J.-P., Avanzi, M., Marcel, S., & Garner, P. N. (2014). Swiss French regional accent identification. In *Odyssey: The speaker and language recognition workshop.*

Leggetter, C. J., & Woodland, P. C. (1994). *Speaker adaptation of HMMs using linear regression* (Tech. Rep. No. CUED/F-INFENG/TR. 181). Trumpington: Cambridge University Engineering Department.

Liang, H., & Dines, J. (2010). An analysis of language mismatch in HMM state mapping-based cross-lingual speaker adaptation. In *Proceedings of Interspeech.* Makuhari, Japan.

Liang, H., Dines, J., & Saheer, L. (2010). A comparison of supervised and unsupervised cross-lingual speaker adaptation approaches for HMM-based speech synthesis. In *ICASSP* (pp. 4598–4601). Dallas, TX, USA.

Liang, H., Hoffmann, S., & Pfister, B. (2014). Capturing speaker-independent prosodic information by syntax tree-based prosody modelling. In *Proceedings of Interspeech.* (under review)

Lodge, A. (1993). *French. From dialect to standard*. London: Routledge.

Métral, J.-P. (1977). Le vocalisme du français en Suisse romande. Considérations phonologiques. *Cahiers Ferdinand de Saussure*, *31*, 145–176.

Morin, Y. C. (2000). Le français de référence et les normes de prononciation. *Cahiers de l'Institut de linguistique de Louvain*, *26*(1), 91–135.

Ney, H. (1999). Speech translation: Coupling of recognition and translation. In *ICASSP* (pp. 517–520). Phoenix, Arizona, USA.

Noth, E., Batliner, A., Kiessling, A., Kompe, R., & Niemann, H. (2000). Verbmobil: the use of prosody in the linguistic components of a speech understanding system. *IEEE Trans. on Audio, Speech and Language Processing*, *8*(5), 519–532.

Paul, D. B., & Baker, J. M. (1992). The design for the Wall Street Journal-based CSR corpus. In *Proceedings IEEE international conference on acoustics, speech and signal processing* (pp. 899–902).

Qian, Y., Wu, Z., Gao, B., & Soong, F. K. (2011). Improved prosody generation by maximizing joint probability of state and longer units. *IEEE Trans. on Audio, Speech and Language Processing*, *19*(6), 1702–1710.

Racine, I., Schwab, S., & Detey, S. (2013). Accent(s) suisse(s) ou standard(s) suisse(s)? Approche perceptive dans quatre régions de Suisse romande. In A. Falkert (Ed.), *La perception des accents du français hors de france* (pp. 41–59). Mons: Éditions CIPA.

Reynolds, D. A., Quatieri, T. F., & Dunn, R. B. (2000). Speaker verification using adapted Gaussian mixture models. *Digital Signal Processing*, *10*, 19–41. Retrieved from http://www.idealibrary.com

Schwab, S., Avanzi, M., Goldman, J.-P., Montchaud, P., Racine, I., et al. (2012). An acoustic study of penultimate accentuation in three varieties of French. In *Proceedings of Speech Prosody.*

Schwab, S., & Racine, I. (2013). Le débit lent des suisses romands: mythe ou réalité? *Journal of French Language Studies*, *23*, 281–295.

Sertling Miller, J. (2007). *Swiss French prosody: intonation, rate, and speaking style in the Vaud canton* (Unpublished doctoral dissertation). Graduate College of the University of Illinois, Urbana-Champaign.

Stan, A., & Giurgiu, M. (2011). A superpositional model applied to f0 parameterization using DCT for text-to-speech synthesis. In *6th conference on speech technology and human-computer dialogue (SpeD)* (pp. 1–6).

Suni, A. S., Aalto, D., Raitio, T., Alku, P., Vainio, M., et al. (2013). Wavelets for intonation modeling in HMM speech synthesis. In *8th ISCA workshop on speech synthesis.* Barcelona.

Wehrli, E. (2007). Fips, a "deep" linguistic multilingual parser. In *Proceedings of the workshop on deep linguistic processing* (pp. 120–127). Stroudsburg, PA, USA: Association for Computational Linguistics.

Wehrli, E., Nerima, L., & Scherrer, Y. (2009). Deep linguistic multilingual translation and bilingual dictionaries. In *Proceedings of the 4th workshop on statistical machine translation* (pp. 90–94). Stroudsburg, PA, USA: Association for Computational Linguistics.

Wester, M., Dines, J., Gibson, M., Liang, H., Wu, Y.-J., Saheer, L., … Yamagishi, J. (2010). Speaker adaptation and the evaluation of speaker similarity in the EMIME speech-to-speech translation project. In *SSW7.* Kyoto, Japan. Retrieved from http://publications.idiap.ch/downloads/papers/2010/Wester_SSW_2010.pdf

Woehrling, C., & de Mareüil, P. B. (2006). Identification of regional accents in French: perception and categorization. In *Interspeech* (pp. 1511–1514).

Wu, Y.-J., Nankaku, Y., & Tokuda, K. (2009). State mapping based method for cross-lingual speaker adaptation in HMM-based speech synthesis. In *Interspeech* (pp. 528–531). Brighton, U.K..

Yamagishi, J., Nose, T., Zen, H., Ling, Z.-H., Toda, T., Tokuda, K., … Renals, S. (2009). A robust speaker-adaptive HMM-based text-to-speech synthesis. *IEEE Trans. on Audio, Speech and Language Processing*, *17*(6), 1208–1230.

Yamagishi, J., Usabaev, B., King, S., Watts, O., Dines, J., Tian, J., … Kurimo, M. (2010). Thousands of voices for HMM-based speech synthesis–analysis and application of TTS systems built on various ASR corpora. *IEEE Trans. on Audio, Speech and Language Processing*, *18*(5), 984–1004.

Young, S., Odell, J., Ollason, D., Valtchev, V., & Woodland, P. (1997). *The HTK book (for HTK version 2.1).* Cambridge: Cambridge University.

Zen, H., & Braunschweiler, N. (2009). Context-dependent additive log f_0 model for HMM-based speech synthesis. In *Interspeech* (pp. 2091–2094).

Zen, H., Nose, T., Yamagishi, J., Sako, S., Masuko, T., Black, A., & Tokuda, K. (2007). The HMM-based speech synthesis system (HTS) version 2.0. In *SSW6* (pp. 294–299).

Zen, H., Tokuda, K., & Black, A. W. (2009). Statistical parametric speech synthesis. *Speech Communication*, *51*(11), 1039–1154.