A TRANSFORMATION-BASED LEARNING APPROACH TO LANGUAGE IDENTIFICATION FOR MIXED-LINGUAL TEXT-TO-SPEECH SYNTHESIS

J.C. Marcadet[1], *V. Fischer*[2], *C. Waast-Richard*[1]

IBM Pervasive Computing, European Voice Technology Development
1) IBM France, 1, place Jean-Baptiste Clément, 93881 Noisy le Grand Cedex, France
2) IBM Deutschland Entwicklung GmbH, Schönaicher Str. 12, D-71032 Böblingen, Germany
{marcadet,waast}@fr.ibm.com, vfischer@de.ibm.com

ABSTRACT

Recent progress in corpus-based concatenative text-to-speech synthesis has generated some interest in systems that are capable of synthesizing text from more than one language. In this paper we describe the language identification component of such a mixed-lingual text-to-speech system. Relying only on the input text, we employ two different methods, namely a transformation based learning approach and a stochastic n-gram approach, and we describe the combination of both methods. While the transformation-based learning approach already produces average error rates of less than 2 percent and outperforms the n-gram classification scheme, the combination of both methods results in a further error reduction of up to 50 percent.

1. INTRODUCTION

Recent progress in corpus based concatenative text-to-speech synthesis has generated some interest in systems that are capable of synthesizing text from more than one language. While multi-lingual speech synthesis focuses on the design of algorithms and architectures that can be used for the construction of synthesizers for a variety of languages [1], it is the goal of mixed-language (or polyglot) speech synthesis to synthesize text from different language within a single sentence [2, 3, 4]. In particular for text-to-speech synthesizers of the second type, automatic language identification is an important prerequisite, since it can free application developers from the sometimes inconvenient need to use markup or annotations, and - more important - it often allows a fast prototyping due to the possible utilization of existing monolingual language resources and algorithms.

Language identification algorithms have been developed for both spoken language and written text input. While spoken language identification is usually based on the phoneme sequence extracted from the speech signal [5], written language identification is frequently solved by the computation of language scores from language specific character n-gram statistics , cf. for example [6, 7]. Although both phoneme sequence and written text are available in a text-to-speech synthesizer, the work presented in this paper concentrates on language identification from written input text, since we believe that this results in a much simpler overall system architecture. In doing so, we employ a "classical" character n-gram method that is augmented by the use of word context information, develop a new, transformation-based learning approach to language identification, and demonstrate the benefits of a combination of both methods.

The remainder of this paper is organized as follows: In Section 2 we give a brief description of the IBM trainable text-to-speech system, which serves as a test-bed for our investigations. We then describe the language identification methods in some more detail (Section 3), before giving experimental results for the various methods in Section 4. Finally, we conclude with a summary and a discussion of future work in Section 5.

2. TTS SYSTEM OVERVIEW

The trainable text-to-speech system developed at IBM serves as a test-bed for our work on mixed-lingual synthesis and is briefly described in this section. A more detailed description can be found, for example, in [8, 9].

Text normalization, text-to-phone-conversion, and phrase boundary generation are performed by a rule-based front-end, and prosody prediction uses decision trees to map a set of features extracted from the front-end to pitch and duration targets for each syllable or phone, respectively. Preprocessed phrases are passed to the back-end search engine, which employs a Viterbi beam search to generate synthetic speech from a set of subphoneme-sized candidate speech segments that are identified by phonetic context decision tree growing during the training of the system.

Since we use a common phonetic alphabet for the construction of our mixed-lingual voices (e.g. English com-

bined with either French, German, or Spanish; cf. [4]), the back-end is fully bilingual, and the search has access to synthesis units from both languages.

In contrast, multiple language dependent front-ends are used, and it is the task of the language identification module described in this paper to tag each input word with the proper language identity. Language tags are interpreted by the synthesizer and cause the corresponding front end in order to create the language specific pronunciation and intonation. The construction of synthetic output speech in multiple languages by the back end search engine includes the removal of inter-phrase silence, which occurs due to the change from one front-end to another; other methods for the smoothing of transitions from one language to the other are subject of current research.

3. LANGUAGE IDENTIFICATION METHODS

While the system architecture sketched in the previous section allows a fast prototyping of voices for arbitrary combinations of languages for which a linguistic front end exists, it limits the sources of information available for language identification to the written input text and therefore sets the general conditions for the two algorithms described in the remainder of this section.

3.1. Transformation Based Learning

Transformation based learning (henceforth TBL) has originally been introduced to tackle the problem of part-of-speech tagging [10, 11], and has since then been applied to several natural language processing tasks including, for example, natural language parsing [12] or machine translation [13].

The main idea of TBL is to start with a simple solution to the problem and iteratively apply transformations that improve the quality of the current solution; in each step the transformation is selected and applied that results in the largest gain according to some pre-defined criterion. The algorithm stops if either the obtained solution does not significantly differ from the previous one, or no more transformations can be selected. Frequently formulated as rewriting rules, transformations are created from a set of rule templates in a training phase that makes use of some reference data, e.g. some carefully annotated input text.

Following the general scheme of the algorithm, our TBL language identification method consists of two components, an initial state annotator and a rule tagger, each of which is described here.

3.1.1. Initial State Annotator

This component is used to assign all possible language tags to a given word. It consists of a lexicon lookup, morphological analysis, unknown word handling, and primary language detection.

Lexicon Lookup and Morphological Analysis. A common dictionary for 5 languages — English (EN), French (FR), German (GR), Italian (IT), and Spanish (ES) — was created from the most frequent words found in language specific lexicons and proper name lists of different size.

language	no. of words
English	54.041
French	329.807
German	350.479
Italian	127.331
Spanish	619.181
common	1.446.942

Table 1. Size of language specific and common dictionary used for TBL based language identification of 5 languages.

In order to reduce the size of the lexicon (cf. Table 1) and to generalize some morphological phenomena, we derived three different kind of morphological rules: Whereas *special character rules* assign language tags based on the occurrence of language specific accented characters (e.g. á $\rightarrow$ ES,ç $\rightarrow$ FR), *suffix and prefix rules* are designed to compress the lexicon by the exploitation of language specific character sequences (e.g. ated $\rightarrow$ EN (suffix), or Schm $\rightarrow$ GR (prefix)). By the creation of 27.166 morphological rules we were able to reduce the lexicon size by more than 85 percent to 214.050 words.

Unknown word handling. Since lexicon and morphological analysis will not cover every single word that can appear in an unknown input text, an attempt is made at this stage to classify unknown words. For that purpose, punctuation marks are tagged with a special tag, capitalized words are lowercased and checked against the lexicon and the morphological rules, and acronyms and numbers are tagged with the most frequent tag in the sentence. The remaining unknown words are labeled with all possible tags.

Primary language detection. Once all input words have been initially labeled, we try to detect the main or *primary* language of the sentence under consideration. For that purpose, we compute the number of tags for each of the languages. Punctuation marks, numbers, acronyms, and capitalized words are not considered in this step, and the primary language is assumed to be those with the largest number of tags.

3.1.2. Rule Tagger

This component reads labeled text produced by the Initial State Annotator and applies contextual rules to reduce the ambiguity of the tags. Lacking a sufficient amount of bilin-

gual training text, we refrained from generating the disambiguation rules by running the TBL training procedure, but manually designed approx. 500 rules. The information uti-

	if	then
1	NEXTWORD = "and"	EN||FR → EN
2	PREVIOUSWORD = "The"	EN||FR → EN
3	PREVIOUSTAG = "EN"	EN||GR → EN

Table 2. Some sample disambiguation rules. (Read the first rule as: "if the next word is "and", then change the tag of the current word from (EN or FR) to EN.").

lized by the rules refers to either the word or the tag in the current position, or to the word or tag to the left or right of the word under consideration; some examples are given in Table 2.

3.2. N-grams with context

The stochastic N-gram approach described in the following was primarily chosen because of its conceptual simplicity and its seamless extendability if more and more languages enter the scenario. However, being entirely corpus-based, and therefore orthogonal to the dictionary-based TBL approach described in the previous section, a combination of both techniques may also be well suited to overcome each individual method's deficiencies.

In a Bayesian approach to word-based language identification, we estimate the language L as

$$\hat{L} \quad = \quad \text{argmax}_L\{P(L|w)\} \tag{1}$$

with maximum a posteriori probability given an input word w. Since it is obviously impossible to compute reliable estimates $P(w|L)$ for all words of a given language, w is rewritten as a sequence of characters, $w = c_1, \ldots, c_n$, and — after taking the logarithm — $P(w|L)$ is approximated by

$$\begin{aligned} Q(w|L) &= \log P(w|L) \\ &= \sum_{i=1}^{n} \log P(c_{i-l}, \ldots, c_i, \ldots, c_{i+r}|L) \end{aligned} \tag{2}$$

Word-based language identification must inevitably fail for homographs, i.e. spellings common to two or more languages. While in a multilingual speech recognition task this problem can be tackled by providing a n-best list of language identities for each vocabulary item [14], the current architecture of our bilingual text-to-speech synthesizer (cf. Section 2) requires a unique identity for each word. Since in many cases the context of the word under consideration can help to resolve ambiguities, we compute the final language score by taking into account the weighted scores of the words w_{i-l}, $l = 1 \ldots n$, and w_{i+r}, $r = 1 \ldots m$, to the left and right of the word w_i to be classified:

$$\begin{aligned} \tilde{Q}(w_i|L) &= \sum_{k=i-n}^{i+m} \alpha_k \cdot Q(w_k|L), \\ & \sum_k \alpha_k = 1, \alpha_k > 0 \end{aligned} \tag{3}$$

and finally select

$$\hat{L} \quad = \quad \text{argmax}_L\{P(L) \cdot \tilde{Q}(L|w)\} \tag{4}$$

as language of w_i. After some initial experiments we decided to consider only the words to the left and right of w_i and fixed the weights to $\alpha_{i-1} = \alpha_{i+1} = 0.1$; at the same time we also experimented with the size of the character n-grams, and found 7-grams ($l = r = 3$, cf. Eqn. (2)) to provide a good compromise between accuracy and memory requirements.

3.3. Combined approach

During the development of both methods described above we found that each algorithm has its own advantages. While the n-gram approach works very well for long words, but has its problems with short words such as prepositions or articles, the TBL method is very successful in classifying the usual words of a language. Therefore, a combination of both approaches may be advantageous, in particular in a bilingual TTS scenario, where we frequently have to deal with "carrier phrases" in a main language that embrace one or more words from a second foreign language.

For that purpose, we combined n-grams and TBL at the output of the Initial State Annotator: If a word is unknown (i.e. neither covered by the reduced dictionary nor by the morphological rules), we use the n-gram method to provide a language tag, rather than tagging it with all labels by default. The so created initial solution is less ambiguous and subject to further disambiguation by the Rule Tagger that uses the original set of rules.

4. EXPERIMENTS

Aiming primarily on a better identification of foreign words in the textual input to our text-to-speech synthesizers we designed test scripts for French, German, and Spanish (FR, GR, and ES, respectively) that contain a reasonable number of words from a second language (usually English proper names or technical terms) as well. Sentences for each test script were extracted from the IBM web pages or from newspapers available on the Web; the characteristics of each script are given in Table 3.

We used an in-house text data base of newspaper arti-

	FR	GR	ES
sentences	50	49	25
English words	195	123	119
French words	1129	0	0
German words	6	795	2
Italian words	5	0	0
Spanish words	8	0	494
punctuation	202	132	99
total	1545	1050	714

Table 3. Number of words from each language in the three different test scripts

cles for the training of n-gram models for English, French, German, Italian, and Spanish. While the total amount of text was approximately the same for all languages (1.2M words) we found significant differences in the number of different words per language, ranging from approx. 53.000 for English to 104.000 for Italian and Spanish; the total number of different words for all five languages was approx. 420.000. For an evaluation of the TBL approach we used the dictionary and morphological rules as described in Section 3.

	n-gram	TBL	combined	gain
FR	6.73	1.36	0.78	42.65
GR	2.86	2.67	1.33	50.19
ES	4.90	1.40	0.84	40.00

Table 4. Word based language identification error rates (in percent) for three different test scripts and methods.

Table 4 reports error rates for both the n-gram method and the TBL method, and compares them to error rates obtained for the combined approach. It becomes evident that the TBL approach outperforms the n-gram method, but that due to the different nature of both methods a very good error rate reduction can be obtained from the combination of both.

5. CONCLUSION

In this paper we have presented the language identification component of a mixed-lingual text-to-speech system. We compared a corpus-based stochastic n-gram approach and a dictionary-based TBL approach and found good gains from a combination of both methods, resulting in a language identification error rates of around 1 percent. Future work must deal with the extension to more languages, the incorporation of more application specific dictionaries, and a further reduction of footprints.

6. REFERENCES

[1] R. Sproat, Ed., *Multilingual Text-to-Speech Synthesis. The Bell Labs Approach*, Kluwer Academic Publishers, Dordrecht, Boston, London, 1998.

[2] C. Traber, K. Huber, K. Nedir, B. Pfister, E. Keller, and B. Zellner, "From Multilingual to Polyglot Speech Synthesis," in *Proc. of the 6th Europ. Conf. on Speech Communication and Technology*, Budapest, 1999.

[3] H. Li, F. Chen, L. Shen, and X. Ma, "Trainable Cantonese/English Dual Language Speech Synthesis System," in *Proc. of the IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, Hong Kong, 2003.

[4] S. Kunzmann, V. Fischer, J. Gonzalez, O. Emam, C. Günther, and E. Janke, "Multilingual Acoustic Models for Speech Recognition and Synthesis," in *Proc. of the IEEE Int. Conference on Acoustics, Speech, and Signal Processing*, Montreal, 2004.

[5] M. Zissman, "Comparison of four approaches to automatic language identification of telephone speech," *IEEE Trans. on Acoustics, Speech and Signal Processing*, vol. ASSP-44, no. 4, pp. 31–44, 1996.

[6] J. Prager, "Linguini: Language Identification for Multilingual Documents," in *Proc. of the 32nd Hawaii Int. Conf. on System Sciences*, Hawaii, 1999, pp. 1–11.

[7] J. Tian, J. Häkkinen, S. Riis, and K. Jensen, "On Text-Based Language Identification for Multilingual Speech Recognition Systems," in *Proc. of the 7th Int. Conf. on Spoken Language Processing*, Denver, 2002.

[8] E. Eide, A. Aaron, R. Bakis, P. Cohen, R. Donovan, W. Hamza, T. Mathes, M. Picheny, M. Polkosky, M. Smith, and M. Viswanathan, "Recent Improvements to the IBM Trainable Speech Synthesis System," in *Proc. of the IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, Hong Kong, 2003.

[9] W. Hamza, R. Bakis, E. Eide, M. Picheny, and J. Pitrelli, "The IBM Expressive Speech Synthesis System," in *Proc. of the 8th Int. Conf. on Spoken Language Processing*, Jeju Island, Korea, 2004.

[10] E. Brill, "A simple rule-based part-of-speech tagger," in *Proc. of the 3rd Conf. on Applied Natural Language Processing*, Trento, Italy, 1992, pp. 152–155.

[11] E. Brill, "Transformation-based error-driven learning and natural language processing: A case study in part-of-speech tagging," *Computational Linguistics*, vol. 21, no. 4, pp. 543–566, 1995.

[12] L. Ramshaw and M. Marcus, "Text Chunking using Transformation-Based Learning," in *Proc. of the 3rd Annual Workshop on Very Large Corpora*, 1995.

[13] S. Cortson-Oliver and M. Gamon, "Combining decision trees and transformation-based learning to correct transferred linguistic representations," in *Proc. of the 9th Machine Translation Summit*, New Orleans, 2003.

[14] J. Tian and J. Suontausta, "Scalable Neural Network based Language Identification from written Text," in *Proc. of the IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, Montreal, Canada, 2004.