Multi-Lingual and Multi-Modal Speech Processing and Applications

Jozef Ivanecky, Julia Fischer, Marion Mast, Siegfried Kunzmann, Thomas Ross, Volker Fischer

IBM Deutschland Entwicklung, AIM Voice Technologies
Schönaicher Str. 220, D-71072 Böblingen
{ivanecky, julifi, mmast, kunzmann, tross, vfischer}@de.ibm.com

Abstract. Over the last decade voice technologies for telephony and embedded solutions became much more mature, resulting in applications providing mobile access to digital information from anywhere. Both a growing demand for voice driven applications in many languages and the need for improved usability and user experience now drives the exploration of multi-lingual speech processing techniques for recognition, synthesis and conversational dialog management. In this overview article we discuss our recent activities on multi-lingual voice technologies and describe the benefits of multi-lingual modeling for the creation of multi-modal mobile and telephony applications.

1. Introduction

Since the mid 90's speech recognition technology made tremendous progress. At that time focus on research was mainly towards speaker dependent, very large vocabulary speech recognition to solve dictation type usage scenarios for a PC [1]. With the introduction of dictation products and their wide distribution to a broad audience research interest moved rapidly on towards speaker independent speech recognition technologies using mainly grammars and small to medium sized vocabularies. In addition, the convergence of mobile phones and embedded devices has driven progress in both research on noise robustness for telephone channels as well as in optimization techniques for very small footprint deployments.

Today, advances in voice technology development and the growing number of information access applications promise an easy and natural access to information in any environment.. Imagine applications like tourist information, traffic jam information, stock quote query systems, and even more voice enabled Internet portals which must deal with content from multiple languages spoken by native, accented or non-native language speakers not only stress pronunciation handling but all levels of the system. Voice solutions might be deployed on small devices like PDA's and navigation systems with a variety of input and output modalities like voice, keyboard, stylus and display, whereas a telephone deals these days with voice input and output only. However, targeting all these rapidly growing solution scenarios requires not only research in the fields of the core technologies, but also on technology standardization of voice user interfaces (VUI), programming languages covering multi-modal input & output, and research on conversational systems that can deal with natural dialogs [2]

Above considerations demonstrate that making multi-language information available anywhere via voice to a large user population requires extensive research in a wide range of multi-lingual voice technologies spanning speech recognition, speech synthesis and conversational dialog management. For efficiency and best system performance it is beneficial to systematically consider and exploit synergies between all employed voice technologies. Herein we will provide an overview on our recent multi-lingual activities which include the definition and use of common phone alphabets for speech recognition and synthesis, acoustic modeling for native and non-native multi-lingual speech recognition, and the development of systems capable of handling directed and conversational dialog in many languages. We also provide insight into functioning prototype development and system view introductions addressing standardization of voice user interface development for consumer devices and telephone systems.

The remainder of the paper is organised as follows: Section 2 provides a brief overview over progress on common phone alphabet definitions and Section 3 describes multi-lingual acoustic modelling experiments including results on non-native speaker recognition. Section 4 provides on overview on our initial activities on speech syntheses dealing with multi-lingual text and exploiting common phone alphabets. Section 5 introduces conversational dialog systems and adaptations for dealing with multi-language speech input. Section 6 focuses on voice user interface standardisation and prototypes examples for client and telephone systems. Finally Section 7 provides a conclusion and some prospect for further work.

2. Evolution of Common Phone Alphabets

The definition of a common phone alphabet for multilingual speech recognition has to deal with at least two conflicting goals: in speech recognition the phonetic inventory of each language should be covered as precise as possible in order to achieve high recognition accuracy, while at the same time as many phones as possible should be shared across languages. Maximizing the overlap will a) efficiently utilize the training data and b) lead to reasonably small acoustic models. A similar tradeoff can be observed for common phone alphabets defined for speech synthesis: while on the one hand the sounds of each language should be kept separate in order to enable high quality synthetic speech for all languages, a less detailed definition may result in a broader variety of individual synthesis units. In particular for small sized segment databases the latter may help to better match the targets requested by the synthesizers linguistic front end, cf. Section 4.

Starting from available, disjoint phonetic alphabets for seven languages (Arabic, British English, French, German, Italian, (Brazilian) Portuguese, and Spanish) which are used within our monolingual speech recognition research activities we have designed two common phonetic alphabets of different detail [3]. In a first step, language specific phone sets were simplified following available SAMPA transcription guidelines (see [4]) which affected each language's phone set to a different degree: While, for example the native French phone set remained unchanged, we gave up syllabic consonants for German, and at the same time introduced new diphthongs for British English. Then, language specific phones mapped to the same SAMPA symbol were merged into a common unit. This resulted in a common phonetic alphabet consisting of 121 phones (65 vowels, 56 consonants) for the seven languages. As can be seen in Table 1, this gave an overall reduction of 60 percent compared to the simplified language specific phonologies.

(a)	total	En	Fr	Gr	It	Es	Pt	Ar
vowels	65	20	17	23	14	10	20	14
consonants	56	24	19	26	32	30	22	29
Total	121	44	36	49	46	40	42	43

Table 1. Number for vowel and consonant phones for seven languages in the detailed common phone set. Languages are British English (En), French (Fr), German (Gr), Italian (It), Spanish (Es), Brazilian Portuguese (Pt), Arabic (Ar).

To increase the overlap we have defined a less detailed common phonetic alphabet, cf. Table 2. We achieved this in three steps: 1) we dropped the distinction between stressed and unstressed vowels for Spanish, Italian, and Portuguese 2) we represented all long vowels as a sequence of two (identical) short vowels and 3) we split diphthongs into their two vowel constituents. In doing so, the average number of languages that contribute to the training data for each of the 76 phones (the sharing factor) increased from 2.28 to 2.53. If we disregard Arabic, the sharing factor increased from 2.74 to 3.56. But this radical inventory reduction caused an increase of the average word error rate by about 7 percent measured on an in-house database if compared to the more detailed common phone alphabet.

(b)	Total	En	Fr	Gr	It	Es	Pt	Ar
Vowels	31	13	15	17	7	5	12	11
Consonants	45	24	19	23	28	24	22	28
Total	76	37	34	40	35	29	34	39

Table 2. Number of vowels and consonants for seven languages in the reduced common phone set.

A further benefit of the reduced phone inventory stems from the fact that additional languages with can be covered with less new phones as with the detailed inventory. The integration of eight additional languages (Table 3) required only 2 additional vowels and 12 consonants which is a result that makes us believe that the slight degradation in accuracy is tolerable and likely to be adjustable by improved acoustic modelling techniques.

	Cz	Jp	Fi	El	Nl	Da	No	Sv
vowels	5	5	8	5	14	14	17	17
consonants	27	23	19	25	22	20	23	24
total	32	28	27	30	36	34	40	41

Table 3. Number of vowels and consonants additional languages integrated into the reduced common phonetic alphabet: Czech (Cz), Japanese (Jp), Finnish (Fi), Greek (El), Dutch (Nl), Danish (Da), Norwegian (No), Swedish (Sv).

3. Multilingual Acoustic Modeling

Multilingual acoustic modeling facilitates the development of speech recognizers for languages with only little available training data, and also allows reduced complexity of application development by the creation of acoustic models that can simultaneously recognize speech from several languages [5]. The use and combination of multilingual acoustic models has also proven advantageous for the recognition of accented speech produced by a wide variety of non-native speakers with different commands of the system's operating language [6].

Acoustic modeling for multilingual speech recognition to a large extend makes use of well established methods for (semi-)continuous Hidden-Markov-Model training. Methods that have been found of particular use in a multilingual setting include, but are not limited to, the use of *multilingual seed HMMs*, the use of *language questions* in phonetic decision tree growing, *polyphone decision tree specialization* for a better coverage of contexts from an unseen target language, and the determination of an appropriate *model complexity* by means of a Bayesian Information Criterion; see, for example, [5, 7] for an overview and further references.

Having now reached a certain maturity, the benefits of multilingual acoustic models are most evident in applications that require both robustness against foreign speakers and the recognition of foreign words. We have simultaneously explored both of these when creating a Finnish name dialer whose application directory consists of a mix of 6,000 Finnish and foreign names, and which is used by native and non-native speakers.

For that purpose, we created acoustic models with different proportions of speech data from Finnish (SpeechDat-II), US-English, UK-English, German, Italian and Spanish. The amount of training material used for the creation of various acoustic models ranges from a mono-lingual Finnish acoustic model created from 70,000 utterances to a multilingual model that was trained from up to 280,000 utterances (approx. 190 hours of speech). As expected, we found a decreasing word error rate when the amount of data increased. Word error rates on the 6,000 foreign names task were between 2.63 percent in case of the monolingual Finnish recognizers and 2.07 percent when the entire data was used for training. More interestingly, we obtained reduced word error rates also when performing digit recognition experiments in Spanish with the so created multilingual acoustic models. These results clearly demonstrate that the acoustic models learn from languages and thus provide robustness for native Spanish speakers.

4. Multilingual Speech Synthesis

Whereas multilingual modeling is an almost well established principle in speech recognition, it is an only emerging concept in the area of speech synthesis, although the need for a better utilization of synergies between both fields has been recently recognized [9]. Despite of some work towards system architectures and algorithms that can be used for the construction of synthesizers for a variety of languages [10,11], today's systems usually achieve speech output in multiple languages by use of two or more language dependent synthesizers (see, for example, [12]), which is frequently accompanied by switching to a different voice.

The IBM trainable text-to-speech system [13] serves as a test bed for our recent work on bilingual, unit selection based speech synthesis, which is briefly sketched in the following. While above mentioned deficiencies are addressed by the construction of a multilingual back end database, in contrast to [14] we do not provide any mixed-lingual text analysis, but employ a set of language specific linguistic front ends in conjunction with a recently developed transformation-based learning approach to language identification [15]. During synthesis, input text is annotated with a language identifier and passed to the corresponding front end that performs text normalization, text-to-phone conversion, and phrase boundary generation. Preprocessed phrases are passed to the back-end that employs a Viterbi beam-search to generate the synthetic speech. The cost function has been revised recently, and now tends to favor long contiguous segments which produces fewer splices and allows preservation of the natural prosody.

The construction of bilingual voices (German/English and Spanish/English) relies on a script of about 10.000 sentences (15 hours of speech, including silence) that include approximately 2000 phonetically balanced sentences as well as a variety of newspaper articles, emails about different topics, proper names, digits and natural numbers, and a number of prompts that are related to popular voice driven applications (e.g. air travel information). While the German or Spanish recording scripts already include some English words, for the construction of bilingual voices each script was augmented by 2000 phonetically balanced English sentences that were read by the same voice talents under the same recording conditions.

For the creation of bilingual Spanish/English and German/English voices, we have utilized common phonetic units to a different degree. For Spanish/English we followed a more conservative design and kept the vowels separate, while all of the consonant phonemes were merged. In contrast, a more aggressive strategy was followed

for the construction of a bilingual German/English synthesizer, where we decided to share not only all consonants between the two languages, but also all vowels, nasal vowels and diphthongs. Multilingual Hidden Markov Models trained with approx. 50.000 utterances from five different languages (English, French, German, Spanish, and Italian) are used to construct the synthesizer's acoustic unit inventory. In order to obtain most accurate alignments, speaker-independent, time synchronous models are transformed into speaker-dependent, pitch synchronous models by running several iterations of the forward-backward algorithm during this process. Phonetic context clustering of the final, pitch-synchronous alignment is used to create a bilingual acoustic decision tree; each leaf of these trees holds a set of subphoneme-sized speech segments from which the output speech is generated.

During the construction of the system, we experienced several advantages from using multilingual models, ranging from the seamless alignment of both the native and the non-native part of the recorded corpus to a better agreement between the speaker's actual pronunciation of foreign words and the dictionary which helps to avoid segmental errors during synthesis. Additional benefits were seen when using bilingual decision trees for the creation of speaker dependent prosody targets. For that purpose, sets of features extracted from each language's front-end are mapped to pitch and duration targets for each syllable or phone, respectively. Training of a common decision tree from both the native and non-native part of the speech database turned out to be an efficient method to overcome data sparseness resulting from the fact that we have recorded only a small amount of non-native data so far. While informal listening tests unveiled no degradation for synthesis in any of the primary languages (German or Spanish) and showed improved synthesis quality for embedded English phrases, the construction of a fully bilingual synthesizer still requires the recording of a larger English corpus.

5. Multilingual Conversational Dialog Systems

The scale of multilinguality for a conversational system can reach from a set of monolingual systems for all languages, where the user chooses in the first utterance the language of choice to be used during the whole dialog, towards completely multilingual systems, where all system components are able to process multilingual input and output.

If built up from several separate monolingual systems, at the beginning of each dialog the system has to decide which language the user prefers. This can be done in different ways:

1. The user can be asked in different languages to choose the preferred language by using touch tones,
2. A language identification module utilizes the first user utterance to determine the language to be used for the remaining dialog,
3. Throughout the whole interaction the utterances can be processed in parallel in each language and the output with the best score determines the language.

The advantage of the latter approach is that a multilingual system can be built with minimal effort, in case systems for different languages exist.

An alternative approach is to combine a multilingual speech recognition system with several monolingual systems for the NLU and dialog that run in parallel. In this case the language of the dialog must be identified from the spoken utterance, either by employing a spoken language identification module, or – in case of grammar based speech recognition – by identifying the language from the grammar that scores best. Drawbacks of this approach are a potentially less accurate speech recognition component, and the possibility of incorrect language identification.

However, the most challenging approach is to create a conversational system in which all components can cope with multilingual input and output. Besides the multilingual speech recognition, multilingual NLU, dialog, answer generation and synthesis component are needed. The advantage is that a user can switch languages between utterances and even within an utterance. The system has to decide for each utterance, in which language the answer will be generated.

There are different aspects to keep in mind, when deciding for one approach or the other. As already discussed, accuracy and application maintenance issues can be an argument to decide for deploying several monolingual language components rather than full multilingual technologies. The parallel language component approach facilitates also the addition of further languages into the system. Another aspect is the practical usage of such a system. In the case of a telephony system, which provides communication in the users' mother tongue, it's rather unlikely that a user will switch between languages within an utterance or even between utterances. When deploying a kiosk system (e.g. at a train station or airport), the demands are quite different. It may be difficult for the system to determine the end of a dialog with one user, and the beginning of the next one. Thus, a system that

allows switching of languages between utterances is a good compromise between flexibility and robust system performance.

An additional issue to consider is the nature of the backend database: If the backend content is available in only one language it needs translation, which is probably not a big deal for dates, prices and numbers, but can be a real challenge for content such as event information (for example the names of performers, venues, streets) which might be language specific and needs special handling during synthesis. In [8] results for two of the above approaches are presented: a system with parallel, monolingual components and a system where all components are bilingual. The developed application covers sport events of the Olympic Games in Athens in 2004. In many respects, the results showed the superiority of the multilingual architecture with parallel language-specific components. Furthermore, the maintenance of this approach turned out to be relatively easy: the system can be constructed from monolingual systems, and adding new languages requires only slight modifications in the overall system. However, the parallel approach is more expensive when it comes to processing power for all parallel components. Results from user tests with the parallel approach were promising and showed that multilingual systems can be built with rather small performance degradation. However, these were only first results, which will be further verified and extended e.g. by adding further languages and with other applications.

6. Multi-Modal Voice User Interfaces

To enable the rapidly growing community of voice application designers to easily develop good voice applications it is required to implement programming models which allow hiding some of the technology complexities especially when multiple modalities for input and output are available. When we started our activities there were no standards for Web-based multi-modal application development established. We thus explored several approaches to support both visual and oral input in a coordinated manner. The multi-modal techniques introduced in this section highlight one of the possible approaches. We picked this particular solution because it allowed natural extension of the VoiceXML model which provides already built-in dialog management capabilities.

In the design of a multi-modal system, one of the important architectural questions is how (and to what extend) user actions and respective application state changes are propagated from one modality to another (i.e. the *synchronization model*). The synchronization model implemented in our browser is a powerful approach to drive applications such as SMS dictation, yet quite compact, and straightforward to be implemented on commercially available PDA such as Compaq iPAQ, Loox and others. The synchronization framework is now introduced in some more detail.

The multi-modal capabilities of this newly developed browser are extensions to the standard VoiceXML input (DTMF and speech recognition) and output (pre-recorded audio or TTS) that support the rendering of HTML pages, and the use of HTML links and forms for user input. We support VoiceXML code to control loading of pages into the HTML component (page-level synchronization) and the HTML component to send specific user-related updates to the VoiceXML component (sub-page-level synchronization):

- Displaying HTML pages is implemented through extensions to the semantics of the VoiceXML <prompt> tag. HTML documents referenced from VoiceXML code are passed to the HTML viewer and treated as a page to be displayed. The page stays displayed until it is rewritten with a new content. Instead of referencing an URL, the designer has an option to construct HTML pages at runtime via ECMA script procedures included as part of the VoiceXML document. Runtime generation of HTML pages is beneficial for multi-modal applications with highly dynamic visual content, which is for example the case in SMS applications.
- The HTML pages displayed can also contain links and forms that can be used to provide explicit synchronization between HTML and VoiceXML components. For example, links with values are used as synchronization anchors that convey the information on user's clicking to the VoiceXML component. Similarly, the values of HTML form variables are propagated to the VoiceXML component upon form completion (on submit). Such sub-page-level synchronization is necessary to implement multi-modal applications that support dynamic filling of HTML forms via a GUI and / or voice.

We refined above mentioned techniques during the development of several multi-modal case study applications. While practical experience from these studies shows that the current set of multi-modal extensions is sufficient for efficient authoring of PDA-scale, multi-modal applications our recent voice user interface activities shifted towards multi-modal design leveraging XHTML and VoiceXML which are widely spread and starting to become a de facto standard.

Very often users have available more than one client (for example, a mobile phone, a PDA, and a notebook) which provide different modalities. Therefore, extending multi-modal concepts to cover multi-client aspects is a

growing focus of interest. The techniques described constitute one of several possible approaches, and were selected because they allow easy extension of the application for a new client setup. Originally, we defined several requirements for the application behavior:

- The application has to cover different clients.
- The incorporation of a new client into the running application has to be maximally simplified.
- The executive code has to be separated from the data as well as dialog control description.

To achieve these requirements we decided to use XML, XSL and XSLT. The dialog flow is described within XSL files. For the different clients the different XSL files are defined wherever required. The dynamic data - like e.g. client side application data, grammars, etc. – is stored in XML files. Beside these two types of files there is also static data like audio files, graphical images, static grammars or other static data which can be used for improving the design for a certain client.

The server side application is waiting for clients' requests. Based on an incoming HTTP request and the type of the client, the executive code selects the appropriate XSL file (which holds the client dependent static data), an XML file (with client independent dynamic data), and additional static data files to generate the requested page (pages). The internal structure of the server side layout (as PHP environment) is depicted in Figure 1.

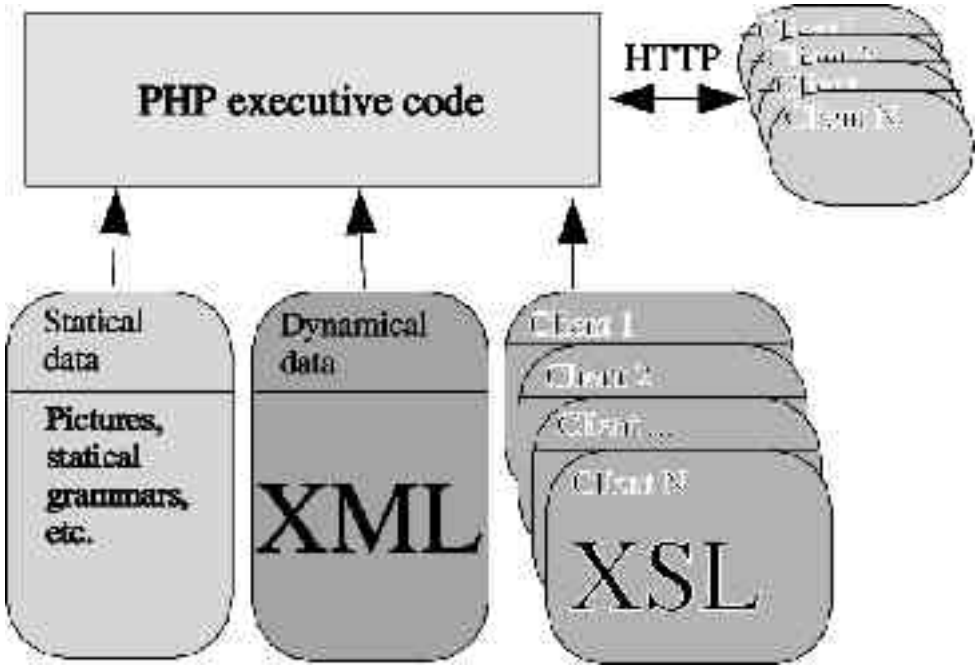

Figure 1: Architecture of Server Side Layout

We refined the above mentioned techniques during the development of several simple multi-client case study applications. Initial experience during this development procedure indicates that this multi-client approach is powerful enough to cover efficiently several different clients.

Above overview on multi-modal design tasks as well as the ability to facilitate voice application development within multi-client scenarios will build the base for our ongoing activities to design multi-lingual applications which leverage recognition, synthesis and conversational understanding technologies.

7. Conclusion and Outlook

Within this article we provided an overview on our recent activities towards multi-lingual speech processing systems. Following a motivation for our overall scope of work we covered the definition and ongoing expansion of a common phone alphabet which builds the base for all our multi-lingual voice technology component activities. Exploiting the various versions of common phone sets we highlighted the progress for multi-lingual speech recognition and synthesis as well as the design and creation of differently architected conversational language understanding applications. To emphasize the importance of overall system and voice user interface

needs we introduced the development of multi-modal and multi-client architectures and its design principles. This evolving infrastructure builds the base of integration of the multi-lingual voice technologies and will ultimately allow mobile access to digital information anywhere. As some of the digital information is very user sensitive or not even available in the user's language, future activities on architecture and technologies may cover the integration of speech biometrics and spoken language translation.

Acknowledgement. The authors wish to make explicit what is obvious to those at home in the community: At some point of our professional career all of us have been associated with the Chair for Pattern Recognition at Friedrich-Alexander-University Erlangen-Nürnberg, headed by Professor H. Niemann. It was Professor Niemann who not only taught us some of the basics of our daily technical work, but also sets the stage for excellence and quality we have been committed to since then. Now, at the occasion of his retirement, we want to say thank you.

8. References

[1] S. Kunzmann, "VoiceType: A Multi-Lingual, Large vocabulary Speech Recognition System for a PC", Proceedings of the 2nd SQEL Workshop on Multi-Lingual Information Retrieval Dialogs, Pilsen, 1997.
[2] S. Kunzmann, "Applied Speech Processing Technologies – our Journey", European Language Resources Association Newsletter, Paris, 2000.
[3] S. Kunzmann, V. Fischer, J. Gonzalez, O. Emam, C. Günther, E. Janke, "Multilingual Acoustic Models for Speech Recognition and Synthesis", Proc. of the IEEE Int. Conf. on Acoustics, Speech, and Signal Processing, Montreal, 2004.
[4] C.J. Wells, "Computer Coded Phonemic Notation of Individual Languages of the European Community", Journal of the International Phonetic Association, Vol. 19, pp. 32-54, 1989.
[5] T. Schultz, A. Waibel, "Language Independent and Language Adaptive Acoustic Modeling for Speech Recognition", Speech Communications, Vol. 35, 2001.
[6] V. Fischer, E. Janke, S. Kunzmann, "Likelihood Combination and Recognition Output Voting for the Decoding of Non-Native Speech with Multilingual HMMs, Proc. of the 7th Int. Conference on Spoken Language Processing, Denver, 2002.
[7] V. Fischer, J. Gonzalez, E. Janke, M. Villani, C. Waast-Richard, "Towards Multilingual Acoustic Modeling for Large Vocabulary Speech Recognition", Proc. of the IEEE Workshop on Multilingual Speech Communications, Kyoto, 2000.
[8] M. Mast, T. Roß, H. Schulz, H. Harrikari, "Different Approaches to Build Multilingual Conversational Systems", Proc. of the 5th International Conference on Text, Speech and Dialogue, Brno, Czech Republic, 2002.
[9] M. Ostendorf, I. Bulyko, "The Impact of Speech Recognition on Speech Synthesis", Proc. of the IEEE 2002 Workshop on Speech Synthesis, Santa Monica, Ca., 2002.
[10] R. Sproat, "Multilingual Text-to-Speech Synthesis. The Bell Labs Approach", Kluwer Academic Publishers, Dordrecht, Boston, London, 1998.
[11] R. Hoffmann, O. Jokisch, D. Hirschfeld, H. Kruschke, U. Kordon, U. Koloska,"A Multilingual TTS System with less than 1 Mbyte Footprint fro Embedded Applications", Proc. of the IEEE Int. Conference on Acoustics, Speech, and Signal Processing, Hong Kong, 2003.
[12] L. Mayfield Tomokiyo, A. Black, K.Lenzo, "Arabic in my Hand: Small-footprint Synthesis of Egyptian Arabic", Proc. of the 8th European Conf. on Speech Communication and Technology, Geneva, 2003.
[13] E. Eide, A. Aaron, R. Bakis, P. Cohen, R. Donovan, W. Hamza, T. Mathes, M. Picheny, M. Polkosky, M. Smith, M. Viswanathan, "Recent Improvements to the IBM Trainable Speech Synthesos System", Proc. of the IEEE Int. Conf. on Acoustics, Speech, and Signal Processing, Hong Kong, 2003.
[14] H. Romsdorfer, B. Pfister, "Multi-Context Rules for Phonological Processing in Polyglott TTS Synthesis", Proc. of the 8th Int. Conf. on Spoken Language Processing, Jeju Island, Korea, 2004.
[15] J.C. Marcadet, V. Fischer, C. Waast-Richard, "A Transformation-Based Learning Approach To Language Identification For Mixed-Lingual Text-To-Speech Synthesis", submitted to: 9th European Conf. on Speech Communication and Technology, Lisbon, 2005.