ISCA Archive
http://www.isca-speech.org/archive

6th International Conference on Spoken Language Processing (ICSLP 2000)
Beijing, China
October 16-20, 2000

A DATA-DRIVEN METHODOLOGY FOR THE PRODUCTION OF MULTILINGUAL CONVERSATIONAL SYSTEMS

Ossama Emam, Jorge Gonzalez, Carsten Günther, Eric Janke, Siegfried Kunzmann, Giulio Maltese, Claire Waast-Richard

IBM Voice Systems
European Speech Research: Cairo, Seville, Heidelberg, Hursley, Rome, Paris

ABSTRACT

This paper describes a data-driven methodology for the design of multilingual conversational systems. The work presented here covers the various aspects of bootstrapping and deploying multilingual systems, such as phone set definition, acoustic modeling, language modeling, and language understanding. For the initial system domain, a Directory Assistance Service has been chosen. Whereas former approaches focused more on the multilinguality of single components of a dialog system we will cover the whole speech-to-speech process of a conversational system.

1. INTRODUCTION

The increasing deployment and acceptance of conversational systems introduces the requirement for multilinguality, particularly in an area such as Europe with its large number of consumers speaking many languages. The system described here is based on the IBM ViaVoice Telephony NLU Toolkit. The architecture is based on the sharing of training data, data structures and software modules across the different languages. Beside an efficient usage of given resources, our design makes it unnecessary to change the architecture of the dialog system with respect to a monolingual systems. This facilitates the addition of new languages, requiring the inclusion of acoustic and linguistic corpora for the new language, but few, if any modifications to the understanding components. As an example application we have chosen a Directory Assistance Service system that allows the access of phone directory information via telephone in a human-like dialog. The user can ask to:

- Get connected to a certain person,
- Get the phone or room number of a person,
- Leave a voice message.

This paper is organized as follows. Section 2 describes the system architecture. Section 3 covers issues of the multilingual phonology, multilingual acoustic models, and multilingual language modeling. The multilingual natural language understanding components are described in section 4. Finally, we summarize our major findings and outline our future work.

2. ARCHITECTURE

The major components of our multilingual (ML) conversational system are a large vocabulary recognition engine, a natural language understanding (NLU) engine, a dialog manager and a text-to-speech (TTS) engine. Figure 1 shows the overall system architecture. Communication between the components of the system is done via a hub. The IVR HUB works as a dispatcher calling and routing information between involved modules. The Telephony Interface handles basic telephony functions, like accepting and disconnecting calls, detection of hang ups, recording and playing back audio and DTMF tone detection. After receiving a call, the audio signal is sent to the ViaVoice Speech Recognition Engine. The speech recognition engine has access to multilingual 8kHz acoustic models and a multilingual language model or grammar. The decoded text is passed to the NLU Control Center. The NLU Control Center controls the communication between the IVR HUB and the NLU modules. The Statistical Classer receives the recognized text and identifies simple concepts. The Canonicalizer assigns canonical values for instances of these basic concepts. These values correspond to data formats required for valid backend requests. The Statistical Parser computes the semantic parse from the classed sentence. The Dialog Manager interprets the parse tree in the dialog context and determines the most suitable backend request. The backend request is issued via the IVR HUB. Additionally, the Dialog Manager generates the system response to prompt back the backend response or to gather additional information from the user if some parameter values are ambiguous or missing to perform a successful backend request. The system response is passed to the ViaVoice TTS engine from where the synthesized speech sound is sent to the Telephony Interface.

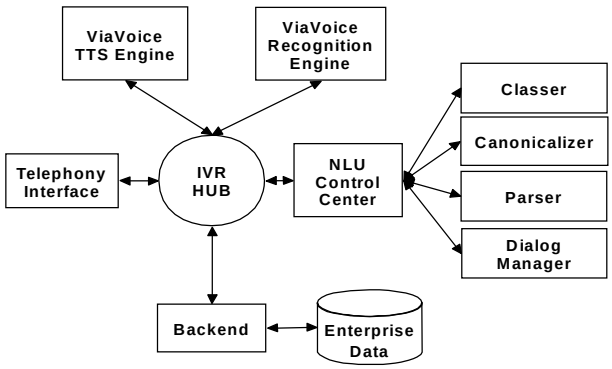

Figure 1: Architecture of the multilingual conversational system.

3. MULTILINGUAL RECOGNITION

The following subsections describe our data-driven approach towards multilingual speech recognizers. We will cover the definition of a common phonology, training ML acoustic models and building ML language models.

3.1. MULTILINGUAL PHONOLOGY

The multilingual recognizer shares acoustic models, language models, and a vocabulary across the respective languages. We have defined a multilingual phone set that covers twelve languages (American English, Arabic, Brazilian Portuguese, British English, Czech, Finnish, French, German, Greek, Italian, Japanese, and Spanish) with some minimum number of phones.

We have developed a general framework that allows us to code word pronunciations in a systematic way while keeping both the accuracy and intuitiveness of the representation at appropriate levels [1].

Our initial work towards the definition of the phone set concentrated in seven languages: Arabic (Ar), Brazilian Portuguese (Pt), British English (En), French (Fr), German (Gr), Italian (It), and Spanish (Es). We started with the already existing alphabets of these languages, which together contain a total of 354 phone models, 132 for vowels or diphthongs, and 224 for consonants. We compared them with SAMPA [2] transcription guidelines and alphabets in order to obtain some language-dependent simplifications and reduce the detail of the phonetic transcription. We reduced the total number of phones to 300, 118 of them for vowels or diphthongs, and 182 for consonants. Details are given in Table 1.

a)	**all**	**En**	**Fr**	**Gr**	**It**	**Es**	**Pt**	**Ar**
vow	132	18	17	23	22	14	24	14
con	224	31	19	37	48	35	24	30
all	356	49	36	60	70	49	48	44
b)	**all**	**En**	**Fr**	**Gr**	**It**	**Es**	**Pt**	**Ar**
vow	118	20	17	23	14	10	20	14
con	182	24	19	26	32	30	22	29
all	300	44	36	49	46	40	42	43

Table 1: Number of vowel (vow) and consonant (con) phonetic units for seven languages: a) original phonetic alphabets; b) simplified alphabets.

In this process, we also assigned IPA phonetic alphabet symbols, through their SAMPA representations, to most language-dependent phonetic units, and merged those that were mapped to the same IPA symbol. The compacted phonetic alphabet so created contains 121 phones, 65 of them for vowels or diphthongs, and 56 for consonants. The overall compression in the number of phonetic units obtained was near one third.

	En	**Fr**	**Gr**	**It**	**Es**	**Pt**	**Ar**
En	-	18	20	20	18	19	19
Fr	2	-	16	17	15	18	15
Gr	8	8	-	21	18	17	18
It	1	5	1	-	20	21	18
Es	1	5	1	10	-	19	17
Pt	1	6	3	9	5	-	16
Ar	3	3	4	1	1	1	-

Table 2: Number of shared consonants (upper right corner) and vowels (lower left corner**)** of the merged alphabet.

Table 2 shows the number of consonant and vowel phones shared by the respective languages. The number of non-shared phones for each language is given in Table 3. The extension of the alphabet so obtained to cover five additional languages (American English, Czech, Finnish, Greek, and Japanese) was straightforward. No additional phones were required for either American English or Finnish. Six consonant phones were added for Czech and Greek, and two more for Japanese. The total number of phonetic units resulted in 129.

	all	**En**	**Fr**	**Gr**	**It**	**Es**	**Pt**	**Ar**
vow	29	9	3	6	-	-	6	6
con	19	-	1	2	8	4	-	8
all	48	9	4	8	8	4	6	14

Table 3: Number of vowel (vow) and consonant (con) phonetic units that belong only to one of the seven languages.

3.2. MULTILINGUAL ACOUSTIC MODELS

The training of a rank-based speech recognition system [3] is a bootstrap procedure that comprises feature extraction, the construction of a set of context dependent, allophonic Hidden Markov Models (HMMs), and the subsequent estimation of the continuous density Gaussian mixture model parameters. Our approach to train ML acoustic models will be outlined based on the language triple British English, French and German.

As a prerequisite, the three languages considered in this subsection use a common acoustic front-end that computes 13 MFCC (including C0) and their first and second order derivative every ten milliseconds. The training data is viterbi-aligned against its transcription in order to obtain a phonetic label for each feature vector. Context dependent HMMs are obtained from the leaves of a decision network [4] that is constructed by asking binary questions about the phonetic context P_i for each feature vector, i=-K,...,K. These questions are of the form: “Is the phone in position i in the subset S_j ?”, and the subsets are derived from meaningful phone classifications commonly used in speech analysis. At each node of the network, the best question is identified by the evaluation of a probabilistic function that measures the homogeneity of the sets of feature vectors that result from a tentative split. Thresholding was used to limit the size of the network. Finally, the data at each leaf of the network is used in a k-means procedure to obtain initial output probabilities whose parameters are then refined by running a several iterations of the forward-backward algorithm.

Using the outlined method, the training of the multilingual acoustic models was performed as follows:

1. After mapping the language dependent phone sets, phonetic context questions and training baseforms to the common phonology, initial monolingual models were trained.

2. The language specific training vocabularies were merged and the monolingual models were used to viterbi-align the training data from each language. Since we noticed that in some cases

the alignment picked a lexeme from a wrong language (e.g. "*a*" is common to English and French), in later experiments we provided each word with a language tag.

3. Based on the (monolingual) alignments a common decision network was constructed to obtain multilingual seed HMMs.

4. Data from all languages was used for the forward-backward refinement of the initial HMM parameters. The latter were obtained by clustering the feature vectors at each terminal node of the decision network by a k-means procedure.

For the first step of the training procedure, outlined above, we used roughly the same amount of training data for each of the three languages (Table 4). This data is a subset of our large vocabulary 22 kHz database down-sampled from 22 to 8 kHz. Then, using the first trilingual model probabilities as prior, we combined different approaches: Maximum a Posteriori (MAP) adaptation [5] followed by a full-build based on the real telephony data. The real telephony data consists into a set of roughly the same amount of training data for each language coming from internal sources and SpeechDat [6]. The resulting systems don't have the same number of context dependent HMMs but always use the same number of elementary mixture densities (30K).

Downsampled data	**En**	**Fr**	**Gr**
N° of speakers	699	1105	500
Training data (hours)	16.2	19.5	19.6
Words (x 1000)	21.6	11.1	18.1
Real telephony data	**En**	**Fr**	**Gr**
N° of speakers	3452	3005	4668
Training data (hours)	76	77	56
Words (x 1000)	26.6	35.5	27.6

Table 4: Acoustic training data

Unlike in [7], language questions are not used in the decision network. We used the phone context, which is presented by the decision network, as the distinguish feature for language identification. In a simple trilingual system, we think language questions are less effective. In addition, this approach required no decoder modifications.

Systems	**En**	**Fr**	**Gr**
MAP Gr, Fr, En	17.2	18.1	23.2
En, Fr, Gr	15.0	17.2	16.2
Fr, Gr	-	16.0	15.7
En, Gr	13.8	-	15.8
En, Fr	12.8	16.1	-
Monolingual	11.5	14.4	15.2

Table 5: Grammar test results

Decoding experiments performed using finite state grammar applications like name dialer requests show (Table 5) a small incremental degradation when more languages are incorporated. It turns out that the MAP approach does not perform as well as the 3 languages full-build, but this could be relative to this restricted test domain and the fact that the latter includes an incremental training.

Further work will deal with approaches like Linear Discriminant Analysis, Bayesian Information Criterion for model complexity determination and speaker clustering methods which have already shown their usefullness in monolingual acoustic modeling.

3.3. MULTILINGUAL LANGUAGE MODELS

For language modeling, we used a combination of trigram-based [8] and of class-based language models [9], trained on multilingual text corpora. For each language a vocabulary was defined. Each word in a vocabulary was tagged with a language tag to minimize erroneous cross-language decoding. For each language we automatically derived word classes using the simulated annealing algorithm [10]. The resulting vocabulary was trained on the corresponding corpus. Class-based language models trained in this way were combined together with static weights. A trigram-based language model, derived by just merging trigrams collected from the entirety of corpora was added to the chain of class-based language models. The language tag of each word prevented mixing up of trigrams coming from different languages.

We can give the following results concerning a multilingual language model trained for British English, Spanish and German. The application domain was the chosen Directory Assistance Service. Vocabulary sizes ranged from 344 to 556 words and the combined vocabulary had 1368 distinct words. The corpus sizes ranged from 0.5 to 1 million of words. The number of classes ranged from 30 to 40 in the class-based language models. A test set was built consisting of 30 sentences uttered by 23 speakers (5 Spaniards, 10 Germans, 8 English). Word error rates ranged from 15% (German) to 25% (British English), while Spanish gave 19.8%.

4. MULTILINGUAL STATISTICAL NLU

4.1. CLASSER, CANONICALIZER, PARSER

Now an overview on the multilingual natural language understanding components will be given. The Classer identifies simple semantic concepts. In the Directory Access System we use the basic concepts: Phone Number, Room Number, and Name. Training from data from the different languages using the same tags to identify the same concept in each language, allows the Classer to identify the concepts irrespective of the language. Only the words that are part of the concepts are give much consideration and so very little language knowledge is required. The Canonicalizer produces standard representations across the languages for the concept values to allow valid backend requests.

Input to the Statistical Parser is the classed sentence. Figure 2 shows the parse tree of example (1) from the Directory Access domain. The Parser, trained from examples in all of the languages, adds tags and labels to the original sentence. These tags and labels are kept common across each of the languages, based on the meaning and concept that is identified by the word. Once each word has its first tag assigned in the tree, there is no language specific information remaining.

Example 1: tell me the phone number for NAME

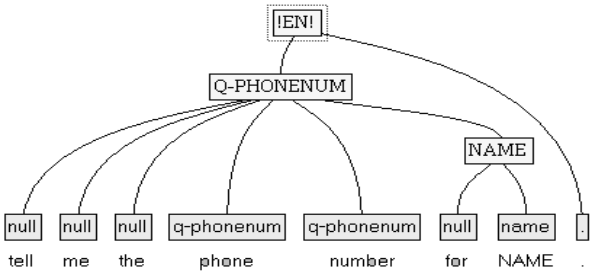

Figure 2: Parse tree of Example 1.

4.2. DIALOG MANAGER, TTS, BACKEND

The FDM (Form-Based Dialog Manager) is a framework for free-flow dialog management [11]. It allows a task oriented, mixed initiative dialog with a user. Each task is modeled as a form. For the Directory Access application 3 forms are used so far: to list a phone or room number of a certain person, to connect to a certain person, and to leave a message. The FDM takes as input the tags and labels from the parse tree and also the canonical values as produces from the Canonicalizer. These values are language independent, the root node of the parse tree identifying the language spoken by the user. The root value allows the FDM to select the response in the correct language for the user.

The system's response is sent to the IBM ViaVoice TTS Engine. It is a formant-based synthesizer and dynamically allows switching of the output language and modifications of prosodic features like f0 and phone duration

The backend has to provide the information needed to perform the task. In the Directory Access application a database with information on around 1000 names is used as backend. The database is accessed across the LAN, with the web server accessing the information using an ODBC interface with SQL requests.

5. CONCLUSION AND FUTURE WORK

The initial Directory Assistant Service system was build for French, German and British English. To add Spanish required simply the addition of Spanish data to the mentioned language dependent components. The addition of further languages is currently underway.

6. ACKNOLEDGEMENTS

We thank our worldwide team working on our multilingual and conversational technology. Special thanks to Francisco Palou, Martin Herzog, Jean-Christophe Marcadet, and Kevin Smith for their contributions to this paper.

7. REFERENCES

1. F. Palou, P. Bravetti, O. Emam, V. Fischer, and E. Janke. Towards a Common Alphabet for Multilingual Speech Recognition Multiple Languages. *In Proc. of the 6th Int. Conf. on Spoken Language Processing*, Beijing, 2000, to appear.

2. J. Wells, W. Barry, M. Grice, A. Fourcin, and D. Gibbon: *Standard Computer Compatible Transcription.* Esprit project 2589 (SAM). Doc. Num. SAM-UCL-037, London, 1992.

3. L. Bahl, de Souza, P. Gopalakrishan, D. Nahamoo, M. Picheny "Robust methods for using context-dependent features and models in a continuous speech recognizer" *Proc. of the IEEE Int. Conference on Acoustics, Speech, and Signal Processing*, Adelaide, 1994

4. L. Bahl and P. de Souza and P. Gopalakrishnan and D. Nahamoo and M. Picheny "Context-dependent Vector Quantization for Continuous Speech Recognition" *Proc. of the IEEE Int. Conference on Acoustics, Speech, and Signal Processing*, Minneapolis, 1993.

5. Gauvain J.L., Lee C.H. "Maximum a-posteriori estimation for multivariate gaussian mixture observations of Markov chain" *IEEE Transactions on Speech and Audio Proc. Vol 2. pp 291-298, 1994.*

6. ELRA European SpeechDat (II) data: http://www.elra.fr

7. Fischer, V., Gonzalez, J., Janke, E., Villani, M., Waast-Richard, C. Towards Multilingual Acoustic Modelling for Large Vocabulary Continuous Speech Recognition. *In Proc. of the MSC2000 Workshop on Multilingual Speech Communications*, Kyoto, 2000, to appear.

8. Jelinek, F., Mercer, R.L.: Interpolated estimation of Markov source parameters from sparse data. *Proceedings of the Workshop on Pattern Recognition in Practice, pp. 381-397*, Amsterdam, May 1980.

9. Brown, P., Della Pietra, V., deSouza, P., Lai, J., Mercer, R.: Class-based *n*-gram models of natural language, *Computational Linguistics, vol. 18, pp. 467-479, 1992.*

10. Jardino, M., Adda, G.: Automatic word classification using simulated annealing, *Proc. of ICASSP 1993, vol. 2, pp. 41-44.*

11. K. A. Papineni, S. Roukos, and R. T. Ward, Free-Flow Dialog Management Using Forms. *Eurospeech 99*, Budapest, 1999.