REAL-TIME SPOKEN LANGUAGE IDENTIFICATION AND RECOGNITION FOR SPEECH-TO-SPEECH TRANSLATION

Daniel Chung Yong Lim[1,2], Ian Lane[1] and Alex Waibel[1]
cdlim@dso.org.sg, ianlane@cs.cmu.edu, waibel.cs.cmu.edu

[1]Language Technologies Institute, Carnegie Mellon University, USA
[2]DSO National Laboratories, Singapore

ABSTRACT

For spoken language systems to effectively operate across multiple languages it is critical to rapidly apply the correct language-specific speech recognition models. Prior approaches consist of either, first identifying the language being spoken and selecting the appropriate language-specific speech recognition engine; or alternatively, performing speech recognition in parallel and selecting the language and recognition hypothesis with maximum likelihood. Both these approaches, however, introduce a significant delay before back-end natural language processing can proceed. In this work, we propose a novel method for joint language identification and speech recognition that can operate in near real-time. The proposed approach compares partial hypotheses generated on-the-fly during decoding and generates a classification decision soon after the first full hypothesis has been generated. When applied within our English-Iraqi speech-to-speech translation system the proposed approach correctly identified the input language with 99.6% accuracy while introducing minimal delay to the end-to-end system.

Index Terms— Language Identification, Speech Recognition, Multilingual Spoken Language Understanding

1. INTRODUCTION

In recent years, voice-enabled human-computer interfaces have become increasingly pervasive in the US. A recent survey [1] cites that 82% of US adults have used a telephone-based automated customer service system in the past year (2009), and for certain tasks users prefered these systems to talking directly with a live customer service agent. In addition to telephone-based spoken dialog systems [2,3], there has been a large growth in the availability of other voice-enabled services. Examples include, systems for automatic transcription of voice messages [4,5], voice-enabled search on mobile devices [6,7], speech-enabled interfaces for self-service Kiosks [8], mobile devices [9] and in-car applications [10].

However, although 18 million adults in the US (8% of the adult population) have limited English proficiency [11], currently deployed systems are limited to a single input language, generally English. If the user is unable to understand or speak this language, it is impossible for them to interact with these systems. The ability to handle multilingual input would extend the capabilities of current systems to support a much wider range of users. This capability would be especially useful for multilingual information kiosks at hospitals, hotels or airport lobbies, speech-to-speech translation systems deployed in multilingual communities, and humanoid robots enabling them to interact and converse in multiple languages.

Creating spoken language systems which can operate across multiple languages, however, remains a challenge. In addition to developing speech recognition and natural language understanding components for each language, the system must also determine which set of language-specific models to use for a given interaction. This can be done by either explicitly eliciting the language required from the user, perhaps by asking them to select it from a menu in a graphical or voice-based user interface, or alternatively, the system must automatically identify the language as the user interacts with the system.

Prior approaches for automatic language identification (LID) and recognition consist of either, first identifying the language being spoken and then applying the appropriate language-specific automatic speech recognition (ASR) engine [12] (LID+ASR); or performing speech recognition in parallel with multiple language-specific engines and selecting the language and recognition hypothesis with maximum likelihood [12] or confidence score [13]. In prior works this approach is called large vocabulary continuous speech recognition (LVCSR)-based LID. Both these approaches, however, introduce a significant delay into the speech processing pipeline. In the LID+ASR approach, the language of the incoming utterance must be determined before speech recognition can begin, thus introducing a delay which could be equally to the length of the input utterance. For LVCSR-based approaches, recognition hypotheses and scores must be generated for all languages before a classification decision can be obtained; this approach is thus limited to the slowest individual ASR

[2] This work was performed when the author was a visiting researcher at the Language Technologies Institute, Carnegie Mellon University.

engine. Even when each individual engine is tuned to operate at near real-time, when a mismatch in language is present between the input speech and models, the time taken to decode the final first-best hypothesis is significantly longer than real-time.

In this work, we propose a novel extension to the LVCSR-based approach to enable language identification and speech recognition to be performed in near real-time. Rather than generating a decision based on the final recognition hypotheses, the proposed approach compares partial hypotheses during decoding. This allows a classification decision to be generated with only a small delay after the first full hypothesis is available. We evaluate the effectiveness of the proposed approach within our English-Iraqi speech-to-speech translation system, and show that it obtains performance similar to that of the traditional approach while introducing minimal delay.

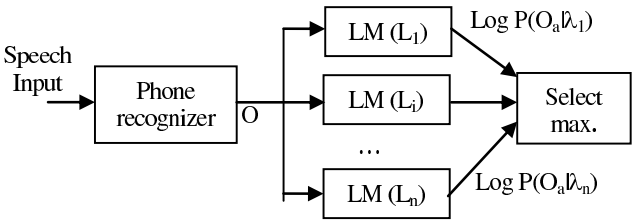
The remainder of the paper is organized as follows. In Section 2 we describe current approaches for phone-based LID, and LID-based multilingual speech recognition. In Section 3 we describe LVCSR-based LID and introduce our proposed real-time LID and recognition scheme. An experimental evaluation for real-time classification of English-Iraqi is described in Section 4. Finally, conclusions are presented in Section 5.

2. PHONE-BASED LANGUAGE IDENTIFICATION FOR ASR MODEL SELECTION

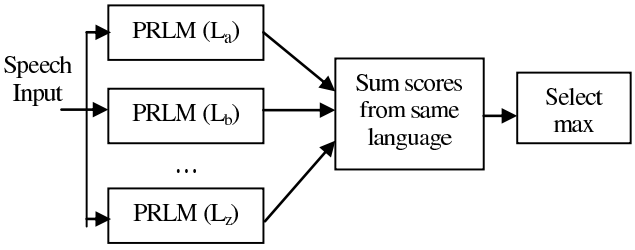
Two popular techniques for spoken language identification are Phone Recognition followed by Language Modeling [14] (PRLM) and Parallel PRLM [15] (PPRLM). In these approaches one (or more) phone recognizers are applied to the input speech and the resulting phone sequence(s) are evaluated using a set of language-specific phone-sequence Markov-models (typically an n-gram phone sequence model). The language of the model(s) with maximum score is selected as the language classification decision. These two approaches are popular for language identification due to their low complexity and significantly lower computational cost compared to LVCSR-based methods. With sufficient computational resources both can operate in a fraction of real-time. As an extension of PPRLM, in [16] we incorporated a CRF-based classifier (PPRLM+CRF), which significantly improved classification accuracy (by up to 25% relative compared to the baseline PPRLM technique). These three approaches are briefly described below.

2.1. Phone Recognition Followed by Language Modeling

Phone Recognition followed by Language Modeling [14] (PRLM) is a common, and computationally low cost method for language identification. A block diagram of a PRLM system is shown in Figure 1a. PRLM applies language-classification models trained on the output of a single phone recognizer and selects the language with maximum model likelihood. During training, utterances for language L_i are converted into phone sequences using a phone recognizer and the output is used to train a phone-language model for that language (LM(L_i)). Models for all n languages $\{L_1, \ldots, L_n\}$ are trained in this manner. During recognition, first, the phone recognizer is applied to the input utterance, and the 1-best phone sequence output (O) is generated. Next, the log likelihood of (O) for each language model is calculated and the language $\hat{l}$ with maximum model likelihood is selected:



(a) Phone Recognition Followed by Language Modeling (PRLM)



(b) Parallel PRLM (PPRLM)

Figure 1: Block diagrams of the phone recognition-based PRLM and PPRLM language identification approaches

$$\hat{l} = argmax_{l \in \{L_1, \ldots, L_n\}} \log P(O|\lambda_l) \quad (1)$$

where $P(O|\lambda_l)$ is the likelihood of the phone language model for language l given for the phone sequence (O).

2.2. Parallel PRLM (PPRLM)

Parallel PRLM (PPRLM) [15] extends on the above approach by applying multiple PRLM systems in parallel. Each system uses a phone recognizer for a different language and the likelihood for language l is summed across the individual systems. A block diagram of this approach is shown in Figure 1b. The training process is similar to that for PRLM, the difference being that multiple sets of language models are trained, one for each phone recognizer. The language $\hat{l}$ with maximum likelihood is:

$$\hat{l} = argmax_{l \in \{L_1, \ldots, L_n\}} \sum_{i=a}^{z} \log P(O_i|\lambda_l) \quad (2)$$

where $P(O_i|\lambda_l)$ is the likelihood of the phone language-

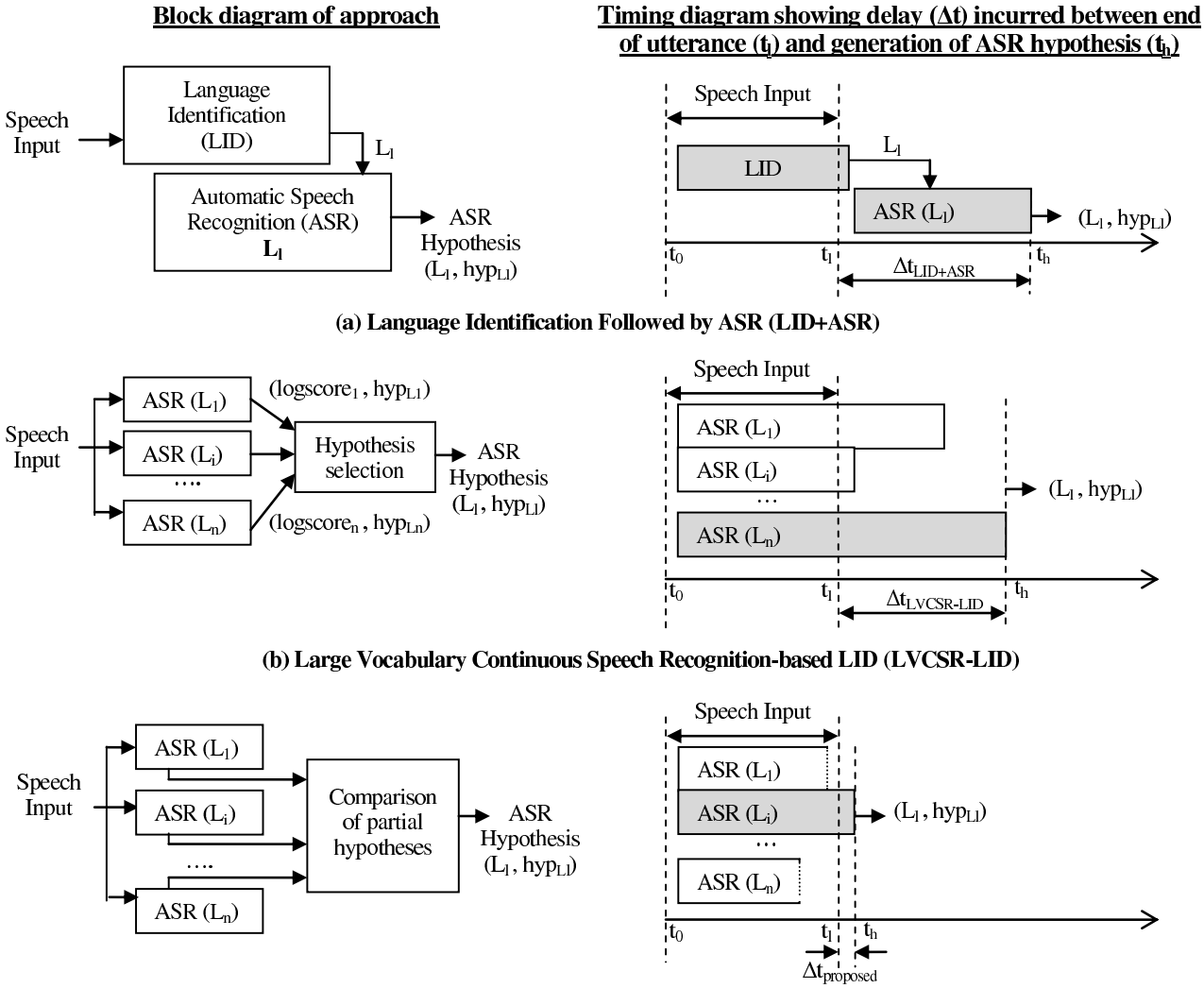


Figure 2: Block and timing diagrams for the three language identification and recognition approaches evaluated

model for language l given the phone sequence from the PRLM system for L_i.

2.3. PPRLM with CRF classifier (PPRLM+CRF)

As an extension to PPRLM, in [16] we proposed a CRF-based approach to incorporate phone confidence scores during language identification. In this approach, first phone sequences from the individual recognizers $\{O_a, \ldots, O_z\}$ were aligned using dynamic programming and a CRF classifier was then applied using phone confidence scores and overall language model likelihood as additional features during classification. The most common label $\hat{l}$ in the output was used as the final classification decision.

2.4. LID-based Model Selection (LID+ASR)

Using one of the LID approaches described above the input utterance can subsequently be recognized using language-specific ASR models. An overview is shown in Figure 2a. With this approach the delay incurred due to language selection, i.e. the delay between the end of the input speech (t_l), until the generation of a recognition hypothesis (t_h), ($\Delta t_{LID+ASR}$) will be t_l, the length of the input utterance. This is assuming the correct language is selected and the ASR-engines operate in real-time. If only the first m seconds of audio are used for classification then:

$$\Delta t_{LID+ASR} = \begin{cases} m & if\ t_l > m \\ t_l & else \end{cases}$$

Using the experimental setup described in Section 4, we determined that for PPRLM+CRF classification no degradation in LID performance was observed when m=10. However, the majority of utterances in the evaluation set had a length < 10sec, so the reduction in delay by incorporating this constraint was small.

3. LVCSR-BASED JOINT LANGUAGE IDENTIFICATION AND RECOGNITION

Compared to traditional LID tasks where only the identity of the language being spoken is required, multilingual spoken language systems also require a hypothesis of what is being said in order to perform spoken language understanding. Therefore, rather than performing LID as a separate pre-selection component, large-vocabulary continuous speech recognition (LVCSR) could be performed in parallel for all languages of interest. The best ASR-hypothesis would be selected identifying both the language ($L_{\hat{l}}$) and recognition hypothesis ($hyp_{L_{\hat{l}}}$) in a single pass. This approach is known as LVCSR-based LID.

3.1. LVCSR-based LID (LVCSR-LID)

Previous works, including [12], have shown that LVCSR-based LID significantly outperforms phone-based approaches. Block and timing diagrams for this approach are shown in Figure 2b. First, speech recognition is performed in parallel using language-specific ASR models for languages $\{L_1, \ldots, L_n\}$. From the resulting set of 1-best ASR-hypotheses $\{(hyp_{L_i}, \text{logscore}_i)\}_{i\in\{1,\ldots,n\}}$ the language $\hat{l}$ with maximum hypothesis likelihood is then selected,

$$\hat{l} = argmax_{l\in\{L_1,\cdots,L_n\}} logscore_l \qquad (3)$$

where $logscore_l$ is the log-scale ASR score of the 1-best ASR hypothesis for language L_l. In this work ASR likelihood scores were used, as no improvement was gained using the normalized confidence described in [13].

For LVCSR-based LID, a 1-best ASR hypothesis for each language is required before the likelihood comparison (3) can be performed. Even when operating on the same multi-core processor the time required to generate hypotheses will vary greatly especially due to the decoding time required when there is a mismatch in language between the input speech and ASR models. The slowest ASR system (shown in gray in Figure 2b) dictates the delay incurred by this approach. In the experimental evaluation in Section 4, $\Delta t_{LVCSR-LID}$ was on average $1.5t_l$ as decoding the input utterance with ASR models of an incorrect language took on average 2.5 x RT.

3.2. Real-time LVCSR-based LID (proposed)

Current LVCSR-based approaches to language identification are unsuitable for real-time applications as ASR hypotheses must be generated for all languages before a classification comparison can be performed. As an extension to this approach we propose comparing the partial hypotheses generated during decoding and halting decoding for less likely languages. Block and timing diagrams for this approach are shown in Figure 2c.

Speech recognition is performed in the same manner as the LVCSR-based approach in 3.1. However, during decoding partial hypotheses are generated for each additional 200ms of input audio. In addition to the ASR score, each partial hypothesis also stores an approximate timestamp (t) for the amount of input audio currently decoded. When decoding is complete the partial hypothesis score and timestamp are replaced with those of the final hypothesis.

During decoding partial hypotheses across languages are compared using a normalized score which takes into account both the ASR likelihood and time lag compared to the fastest system. First, the log-scale ASR score is normalized to compensate for the difference in the amount of audio decoded and the log likelihood ratio compared to the fastest system is calculated. A time-lag penalty is then applied to penalize languages for which decoding is not proceeding. For language l the normalized partial hypothesis score is:

$$score_l = \frac{logscore_l}{logscore_{max}} \cdot \frac{t_{max}}{t_{max} - t_l} - w_{penalty} \left(\frac{t_{max} - t_l}{t_{max}}\right)$$

where:

$$t_{max} = argmax_{l\in\{L_1,\cdots,L_n\}} t_l$$

$logscore_{max}$ is log-scale ASR score for the system with t_{max}, t_l is the amount of audio processed by language l's decoder and $w_{penalty}$ is the time-lag penalty.

When the partial hypothesis for language l is updated its score ($score_l$) is compared to a predefined threshold $\emptyset$ and decoding is halted for languages where the hypothesis falls below this threshold. In the experimental evaluation described in this work a threshold ($\emptyset$) of 0.8 was used and a time-lag penalty ($w_{penalty}$) of 0.5. The performance of the proposed approach was not sensitive to the exact settings of these parameters.

4. EXPERIMENTAL EVALUATION

The classification accuracy and processing delay incurred by the joint LID and recognition approach proposed in Section 3.2 was evaluated within the CMU-TransTAC English-Iraqi Speech-to-Speech translation system [18]. Within this system the proposed approach simplified the user interface, enabling parties of users to interact using a single microphone and single push-to-talk button. Users push and hold down the button while speaking into the microphone and LID is used to select the appropriate translation direction. Language classification accuracy was evaluated on the two standard TransTAC test-sets shown in Table 1. The classification accuracy obtained using the three phone-based approaches described in Section 2 and the two LVCSR-based methods introduced in Section 3 are shown in Figure 3.

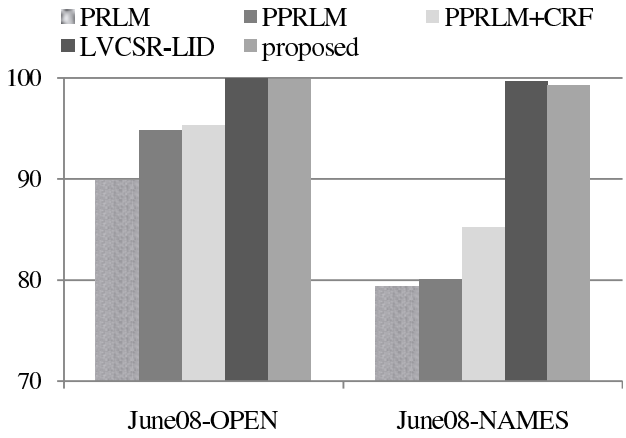


Figure 3: Classification accuracies for phone-based and LVCSR-based language identification in identifying English and Iraqi-Arabic

Table 1: Overview of Training and Evaluation Corpora ([1] GlobalPhone Corpora [17], [2] TransTAC Program Data)

	English	Iraqi/Arabic
Training Corpora		
Phone Recog[1]	20hrs	20hrs
AM (ASR)[2]	150hrs	350 hrs
LM (ASR)[2]	3M words	3M words
LID Models[2]	4 hrs	4 hrs
Evaluation Corpora		
June08-OPEN[2]	54 mins	54 mins
June08-NAMES[2]	8 mins	16 mins

3.1. Comparison of LID performance

First, the language classification accuracy of the three phone-based approaches described in Section 2 was evaluated. For the PRLM system a single English phone recognizer was used, and for the PPRLM and PPRLM+CRF approaches Arabic and English phone recognizers were applied. The three phone-based LID systems were trained on 8 hours of speech, split 50/50 between English and Iraqi. The performance in terms of language classification accuracy is shown in Figure 3. On both test-sets the PPRLM+CRF system obtained higher classification accuracy than the PRLM and PPRLM approaches. On the June08-OPEN set an accuracy of 95.3% was obtain and 85.6% was obtained on the June08-NAMES set. Using this approach a language misclassification occurred for 1 in 10 to 1 in 20 utterances, which is much too high for a fieldable system.

Next, the performance of the LVCSR-LID approach described in Section 3.1 was evaluated. Recognition was performed in parallel with the English and Iraqi speech recognition systems described in Table 2. After recognition completed, the language of the hypothesis with maximum ASR score was selected as the output decision. Compared to the phone-based approaches, LVCSR-based LID obtained

Table 2: Overview of English and Iraqi ASR Systems

	English	Iraqi/Arabic
Model Details		
Codebooks	3000	6000
Max. Gaussians	32	64
Training Method	ML	boostedMMI [19]
Recognition Accuracy WER (RTF)		
June08-OPEN	20.7% (1.0)	27.6% (1.2)
June08-NAMES	26.9% (1.2)	34.5% (1.1)

Table 3: Average per-utterance delay (in milliseconds) from the end of speech input until a recognition hypothesis is available

	June08-OPEN (E/I)	June08-NAMES (E/I)	Avg.
LID+ASR	5038 / 7303	2017 / 4000	4590
LVCSR-LID	8099 /10511	2748 / 5033	6800
proposed	386 / 1575	493 / 507	740
manual	130 / 1317	290 / 346	520

significantly higher classification accuracy: 99.9% for the June08-OPEN set and 99.6% for June08-NAMES.

Finally, our proposed approach, where language identification is performed using partial-hypotheses was evaluated. Using this approach the classification accuracy was comparable to that obtained in LVCSR-LID. Classification accuracies of 99.9% and 99.4% were obtained on the June08-OPEN and June08-NAMES sets respectively. Interestingly, when hypotheses scores were discarded and language identification was selected purely based on time-lag the degradation compared to the proposed approach was small. When ASR scores were not considered classification accuracies of 99.8% and 98.9% were obtained for the two test-sets.

4.2. Comparison of Incurred Delay

In addition to classification accuracy, processing speed and more importantly the delay introduced into the end-to-end system is a critical measure of usability. Using the systems evaluated in Section 4.1, we calculated the average per-utterance delay from the end of speech input until a recognition hypothesis was available. For the LID+ASR approach, PPRLM+CRF classification was applied. Evaluation was performed on a desktop PC with a 2.8GHz quad-core Intel7 processor and 8GB memory. The ASR systems used in this work had been designed to operate at near real-time. The English system operated with an average real-time factor (RTF) of 1.1, and the Iraqi system operated with an RTF of 1.2. The delays incurred by each approach and for the manual case (i.e. when the language is pre-selected by the user) are shown in Table 3.

Both the LID+ASR method (described in Section 2.4) and the LVCSR-based approach (Section 3.1) introduce significant delays into the end-to-end system, on average 5 and 7 seconds per-utterance respectively. The proposed

approach however, introduces little delay showing the effectiveness of the proposed approach.

5. CONCLUSIONS AND FUTURE WORK

In this work we developed a novel method for joint language identification and speech recognition that can operate in near real-time. The proposed approach compares partial hypotheses generated during decoding and obtains a language identification decision soon after the first full hypothesis has been generated. The proposed approach obtains similar classification accuracy to a LVCSR-based language identification (LID) system while operating in near real-time. When applied within our English-Iraqi speech-to-speech translation system the proposed approach correctly identified the input language with 99.6% accuracy while introducing no additional delay to the end-to-end system.

In future work we intend to evaluate the proposed approach for language classification and recognition across a large number of languages. We also intend to extend it to operate across a cloud computing setting where recognition may be performed on heterogeneous servers.

5. ACKNOWLEDGMENTS

We would like to thank Qin Jin (from InterACT labs, Carnegie Mellon University) for providing the English and Arabic phone-recognizers used in this work and Matthias Paulik and Roger Hsiao (also from InterACT labs) who provided the TransTAC English and Iraqi speech recognition engines. This work was in part supported by the US DARPA under the TransTAC (Spoken Language Communication and Translation System for Tactical Use) program. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of DARPA.

6. REFERENCES

[1] Forrester Consulting, "Driving Consumer Engagement With Automated Telephone Customer Service", October 2009.

[2] J. Williams, "Demonstration of a POMDP voice dialer," in Proc Demonstration Session ACL-HLT, 2008.

[3] A. Raux, D. Bohus, B. Langner, A. Black, M. Eskenazi, "Doing Research on a Deployed Spoken Dialogue System: One Year of Let's Go! Experience," In Proc. Interspeech, 2006

[4] Google Voice Voicemail Transcription Service, (http://www.google.com/mobile/apple/app.html)

[5] SpinVox Voicemail Transcription Service, (http://www.spinvox.com)

[6] Google Mobile App with Voice Search, (http://www.google.com/mobile/apple/app.html)

[7] Yahoo OneSearch with Voice, (http://mobile.yahoo.com/onesearch/voice)

[8] A. Karpov, A. Ronzhin, "An Information Enquiry Kiosk with a Multimodal User Interface," In Proc. PRIA-2008, pp. 265-268, 2008

[9] G. DiFabbrizio, T. Okken, and J. Wilpon, "A Speech Mashup Framework for Multimodal Mobile Services," In Proc. ICMI-MLMI, pp. 1-10, 2009.

[10] H. Pon-Barry1, F. Weng and S. Varges, "Evaluation of Content Presentation Strategies for an In-car Spoken Dialogue System," In Proc. Interspeech, pp. 1930-1933, 2006

[11] R. Capps and J. Passel, "Describing Immigrant Communities", Immigration Studies Program, The Urban Institute, December 2004

[12] T. Schultz, I. Rogina, and A. Waibel, "LVCSR-based Language Identification", In Proc. ICASSP, pp. 1827–1830, 1996

[13] F. Metze, T. Kemp, T. Schaaf, T. Schultz, and H. Soltau, "Confidence measure based Language Identification", In Proc. ICASSP, pp. 1827–1830, 2000

[14] T. Hazen and V. Zue, "Automatic language identification using a segment-based approach", In Proc. Eurospeech, pp. 1303-1306, 1993

[15] M. Zissman and E. Singer, "Automatic language identification of telephone speech messages using phoneme recognition and n-gram modeling", In Proc. ICASSP, pp. 305-308, 1993

[16] D. Lim and I. Lane, "Language identification for speech-to-speech translation.", In Proc. Interspeech, pp. 204-207, 2009

[17] T. Schultz and A. Waibel, "Language independent and Language Adaptive Acoustic Modeling for Speech Recognition", in Speech Communication, August 2001, pp. 901-904

[18] N. Bach, R. Hsiao, M. Eck, P. Charoenpornsawat, S. Vogel, T. Schultz, I. Lane, A. Waibel and A. Black, "Incremental Adaptation of Speech-to-Speech Translation.", In Proc. HLT-NAACL, pp. 149–152, 2009

[19] D. Povey, D. Kanevsky, B. Kingsbury, B. vana Ramabhadran, G. Saon, and K. Visweswariah, "Boosted MMI for model and feature-space discriminative training," In Proc. of ICASSP, 2009