# NineOneOne: Recognizing and Classifying Speech for Handling Minority Language Emergency Calls

*Udhyakumar Nallasamy, Alan W Black, Tanja Schultz and Robert Frederking*

Language Technologies Institute,
School of Computer Science, Carnegie Mellon University, Pittsburgh, PA 15213, USA
ref@cs.cmu.edu

In this paper, we describe NineOneOne (9-1-1), a system designed to recognize and translate Spanish emergency calls for better dispatching. We analyze the research challenges in adapting speech translation technology to 9-1-1 domain. We report our initial research towards building the system and the results of our initial experiments.

## 1. Introduction

In many U.S. localities, when emergency calls are received in languages other than English (primarily Spanish) the dispatching center connects the call to the Language Line human translation service [1] to translate for them. Though using human translators assisting callers during emergency calls might seem optimal, this scheme actually has serious shortcomings. The process is very slow, especially in starting up, and the translators are unfamiliar with the task, resulting in very poor quality service.

We are applying and adapting speech translation technology to the domain of 9-1-1 emergency dispatching. (The standard emergency telephone number in the United States is "9-1-1".) The 9-1-1 domain has many research challenges, but we believe it is also a feasible domain for a real-world speech translation application. The domain is challenging because it requires real-time operation and the recognition and translation of stressed telephone-quality speech in multiple dialects; it is still feasible because we have in-domain data, there are strong vocabulary and task constraints, and perfection is not required. This domain also has clear, significant social value, addressing the chronic shortage of translation of Spanish (and in larger cities, a large number of other languages) at U.S. emergency dispatch centers. While we currently are working on Spanish/English, the approaches we use are largely language-independent.

We expect that this project will eventually lead to an actually deployed system that performs full automatic speech recognition (ASR) and targeted, classification-based machine translation (MT). This would be an important scientific result, validating for the first time the hypothesis that speech can be adequately recognized and translated for real-world use. Our initial work is aimed at demonstrating that we can produce ASR and utterance classification of sufficient quality to allow the development of such a practical, limited-domain system. The results reported here achieve that goal.

## 2. Domain Overview

When someone dials 9-1-1 in most places in the United States, they are connected to a special dispatching center. The operators there have been trained to perform a rapid triage, or categorization, of emergency calls. The major initial decision is whether to send police, fire, or medical personnel; the appropriate units are dispatched as soon as this decision is made. While the responding unit is en route, the dispatcher attempts to elicit more details about the emergency from the caller. The additional details are intended to help the responding unit prepare for the situation that they will encounter. For example, the police want to know the level of violence involved in advance. Will they be walking into an armed confrontation, or is someone reporting a crime that occurred yesterday? Another important issue is the exact location of the emergency; although the 9-1-1 equipment displays the phone number and address of the call, the information is not always correct, the emergency may not be at the location of the telephone, and cellular telephones (and Voice-over-IP) do not generally provide 9-1-1 location information. As the dispatcher elicits more information about the emergency, they radio it to the responding unit.

## 3. System Architecture

Full reliable speech-to-speech translation is still beyond our capabilities, especially for real-time human-directed conversation. However in many applications, full translation is not actually required, and a more limited form is adequate [2, 3]. Based on the domain's characteristics, we are following a highly asymmetrical design for the eventual full system [4], see Figure 1. The dispatcher is already seated at a workstation, and we intend to keep them "in the loop", for both technical and social reasons. In the dispatcher-to-caller direction, we can work with text and menus, so we require

- ***no*** English ASR,
- ***no*** true English-to-Spanish MT, and
- simple, domain-limited, Spanish speech synthesis.

The caller-to-dispatcher direction is much more interesting. In this direction we require

- Spanish ASR that can handle emotional spontaneous telephone speech in mixed dialects,
- Spanish-to-English MT, but
- ***no*** English Speech Synthesis.

In the final system, we will adapt an approach that was demonstrated in the joint NSF/EU "Nespole!" project [5]: Domain Action (DA) classification [6, 7]. We use the term DA to refer to the combination of a general Speech Act with domain-specific concepts. DAs capture *speaker intention* in a limited-domain system, rather than detailed literal meaning, and thus help the MT system cope with ungrammaticality and ASR errors. Example DAs in this domain might be Request-Ambulance ("I need an ambulance!") or Giving-Address ("I live at 2635 Rodeo Drive."). Once we have classified the utterance into a DA, the next step in the eventual full system will be to identify and translate just the arguments of the DA. An example output argument structure for the second DA above might be (address-number="2635", address-street="Rodeo Drive).

In this initial work, we only seek to demonstrate that we can carry out utterance classification on ASR output of sufficient quality to support this approach to MT. The actual MT system will be developed in a follow-on project.

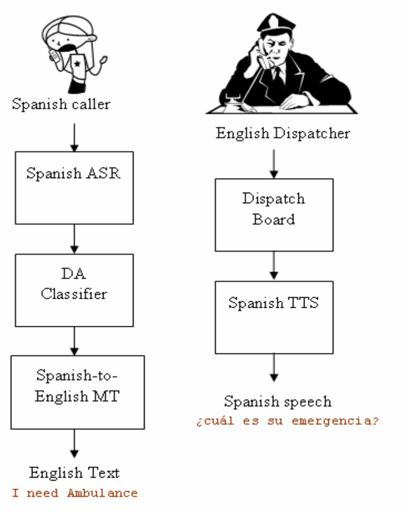

Fig.1 System Flow: a) Caller to Dispatcher and b) Dispatcher to Caller

## 4. Automatic Speech Recognizer (ASR)

The Spanish ASR system is built using the Janus Recognition Toolkit (JRTk) [8] featuring the HMM-based IBIS decoder [9]. The acoustic models are trained on 50 Spanish 9-1-1 calls, which amounted to 4 hours of speech data. The system uses three-state, left-to-right, sub-phonetically tied semi-continuous models with 400 context-dependent distributions with the same number of codebooks. Each codebook has 32 gaussians per state. The front-end feature extraction uses standard 32 dimensional Mel-scale cepstral coefficients and applies Linear Discriminant Analysis (LDA) calculated from the training data. The vocabulary size is 65K words. The language model consists of a trigram model trained on the manual transcriptions of 40 calls and interpolated with a background model trained on GlobalPhone Spanish text data consisting of 1.5 million words [10]. The interpolation weights are determined using the transcriptions of 10 calls. The test data consists of 15 telephone calls from different speakers, which amounts to a total of 1 hour. The perplexity of the test set according to the language model is 96.7. The accuracy of the Spanish ASR on the test set is 76.5%.

## 5. Utterance Classification

As described in Section 3 above, the speech recognizer output needs to be classified into domain-specific DAs for eventual translation, as well as dialogue control. So as an initial evaluation, experiments were carried out on classifying manual audio transcriptions of the 9-1-1 calls used to train the speech recognizer. We extracted all the dialog turns in the transcriptions between the human translator and the caller and labeled them according to the DA of the caller. The following tags were used to label each turn of interaction between the interpreter and the caller: Giving-Name, Giving-Address, Giving-Telephone-Number, Requesting-Medical-Assistance, Requesting-Fire-Service, Requesting-Police, Reporting-urgency-or-injury, Yes, and No. The tags "Yes" and "No" are labels for the answers to yes-no questions. An additional tag called "Others" is used to label all other cases. The database has a total of 845 labeled turns.

We used the Support Vector Machines (SVM) [11] implementation in the WEKA Machine learning toolkit [12] to classify the turns based on Domain Acts. Simple bag-of-words binary features are used for classification. Individual accuracies for each tag are given in Table 1 below.

| Tag | Frequency | Accuracy (%) |
|---|---|---|
| Giving Name | 80 | 57.50 |
| Giving Address | 118 | 38.98 |
| Giving Phone number | 29 | 48.28 |
| Req. Ambulance | 8 | 62.50 |
| Req. Fire Service | 11 | 54.55 |
| Req. Police | 24 | 41.67 |
| Reporting Injury/Urgency | 61 | 39.34 |
| Yes | 119 | 52.94 |
| No | 24 | 54.17 |
| Others | 371 | 75.74 |

Table 1: Utterance Classification Accuracy for all tags

The overall accuracy of the classifier on 10-fold cross-validation is 60.12%. However, if we leave out all mis-classifications of tags as "Others", the accuracy of the classifier improves to 68.8%. This is interesting due to eventual dialogue system design issues. If the system classifies an utterance with the "Others" tag, it can prompt the caller with more specific questions (e.g. "Do you need an ambulance?") to understand their intent. Thus the accuracy without "Others" more accurately reflects the expected performance of the eventual full system. We are currently experimenting with more complex classifier features such as language model perplexity, word classes, etc. In the full paper, we will also present results on classifying ASR output, rather than manual transcriptions.

## 6. Conclusion

The work reported here demonstrates that we can produce Spanish ASR for Spanish emergency calls with reasonable accuracy (76.5%), and classify manual transcriptions of these calls with reasonable accuracy (68.8% ignoring the "Others" category). These results are clearly good enough to justify the next phase of research towards an eventual deployable prototype system.

## 7. Acknowledgements

This project is funded by NSF Grant No: IIS-0627957 "NineOneOne: Exploratory Research on Recognizing Non-English Speech for Emergency Triage in Disaster Response". Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of sponsors.

## 8. References

[1] http://www.languageline.com

[2] D Stallard, F Choi, C Kao, K Krstovski, P Natarajan, R Prasad, S Saleem, and S Subramanian, The BBN 2007 Displayless English/Iraqi Speech-to-Speech Translation System', In Proc Interspeech, Belgium, 2007

[3] Y Gao, B Zhou, R Sarikaya, M Afify, H Kuo, W Zhu, Y Deng, C Prosser, W Zhang, and L Besacier, IBM MASTOR SYSTEM: Multilingual Automatic Speech-to-Speech Translator, In Proc First International Workshop on Medical Speech Translation, ACL, USA, 2006

[4] R Frederking, A Rudnicky, C Hogan, and K Lenzo, Interactive Speech Translation in the Diplomat Project, Machine Translation Journal 15(1-2), Special issue on Spoken Language Translation, June 2000

[5] A Lavie, F Balducci, P Coletti, C Langley, G Lazzari, F Pianesi, L Taddei, and A Waibel, Architecture and Design Considerations in NESPOLE!: a Speech Translation System for E-Commerce Applications. In Proc HLT, pp 31-34, USA, March 2001

[6] L Levin, C Langley, A Lavie, D Gates, D Wallace, and K Peterson. Domain Specific Speech Acts for Spoken Language Translation, In Proc SIGDIAL, Japan, July 2003

[7] C Langley, Domain Action Classification and Argument Parsing for Interlingua-Based Spoken Language Translation. PhD thesis, Carnegie Mellon University, Pittsburgh, PA, 2003

[8] M Finke, P Geutner, H Hild, T Kemp, K Ries, and M Westphal, The Karlsruhe-Verbmobil Speech Recognition Engine, In Proc ICASSP, pp. 83-86, Germany, 1997

[9] H Soltau, F Metze, C F¨ugen, and A Waibel, A One Pass-Decoder Based on Polymorphic Linguistic Context Assignment, In Proc ASRU, Italy, December 2001

[10] T Schultz, M Westphal, and A Waibel, The GlobalPhone Project: Multilingual LVCSR with JANUS-3, In Proc. Multilingual Information Retrieval Dialogs: 2nd SQEL Workshop, pp 20-27, Czech Republic, April 1997

[11] C J C Burges, A tutorial on support vector machines for pattern recognition, In Proc Data Mining and Knowledge Discovery, pp 2(2):955-974, 1998

[12] S R Garner, WEKA: The Waikato environment for knowledge analysis, In Proc New Zealand Computer Science Research Students Conference, pp 57-64, New Zealand, 1995