Parallel combination of speech streams for improved ASR

João Miranda[1,2]*, João P. Neto*[1]*, Alan W Black*[2]

[1]INESC-ID / Instituto Superior Técnico, Lisboa, Portugal
[2]School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, USA
jrsm@l2f.inesc-id.pt, Joao.Neto@inesc-id.pt, awb@cs.cmu.edu

Abstract

In a growing number of applications, such as simultaneous interpretation, audio or text may be available conveying the same information in different languages. These different views contain redundant information that can be explored to enhance the performance of speech and language processing applications. We propose a method that directly integrates ASR word graphs or lattices and phrase tables from an SMT system to combine such parallel speech data and improve ASR performance. We apply this technique to speeches from four European Parliament committees and obtain a 16.6% relative improvement (20.8% after a second iteration) in WER, when Portuguese and Spanish interpreted versions are combined with the original English speeches. Our results indicate that further improvements may be possible by including additional languages.

Index Terms: multistream combination, speech recognition, machine translation

1. Introduction

Our main goal is to explore, in the area of speech-to-speech translation, the frontier between ASR and MT, and how each other benefits with their integration. In several applications, the same or similar information is conveyed by speech in different languages. Probably the two best applications are simultaneous interpretation at the United Nations and the European Parliament. In the latter, committee meetings and plenary sessions are interpreted into each of the languages of the 27 Member States, but are not always transcribed manually, and when they are, there is usually a significant delay between the time when the speech is produced and when the transcription is produced. The use of state-of-the-art ASR systems to perform this transcription automatically is relatively error-prone, given that many speakers choose to speak in a non-native language (usually English), and that the interpreted speeches are full of disfluencies (repairs and filled pauses) . Other instances where multiple versions of the same speech in different languages are available include lectures, broadcasts of sport events, or TV shows.

Most attempts at combining ASR and MT models have traditionally been focused on a sequential combination of these models, for speech-to-speech or speech-to-text translation. However, a number of authors have sought to combine ASR and MT in a parallel fashion. Some of these methods are used to combine speech with a text stream, usually for an application such as machine-aided human translation [2, 3], although a few works have considered combining multiple speech streams [4, 5].

The aim of this work is, therefore, to propose a method that combines the information of recognizers in different languages to yield improved recognition results. In order to connect the language pairs, we use phrase tables trained for a Statistical Machine Translation (SMT) [1] system. These use, as one of their knowledge sources, a phrase table consisting of pairs of the form (source phrase, target phrase). Using these phrase pairs, it is possible to find potential correspondences between speech in different languages, as described in Section 2. The speech recognition models are then biased towards these correspondences. Since this agrees with the knowledge that the many speech versions represent the same message, we expect it to improve recognition performance. Our method differs from previous work [4, 5] in that it does not directly depend on a translation engine decoder, but rather finds phrase pairs that appear both in a phrase table and in lattice pairs. Using lattices instead of N-best lists or 1-best recognizer outputs allows for greater flexibility in the selection of different hypotheses. Therefore, we are able to cope in a more effective way with simultaneously interpreted speech, or with lower resource languages, two scenarios that lead to a significant reduction in recognition performance. Also, no prior alignment between speech segments is required, since one is implicitly constructed during the execution of our algorithm. Our method also scales to an arbitrary number of languages, in cases in which information in more than two languages is available. Finally, through the tuning of a maximum delay parameter δ, it is possible to integrate the improvements presented here into a real-time system.

The rest of this paper is organized as follows. Section 2 details the proposed method. Section 3 describes the experiments that assess the improvements to our system. Finally, section 4 draws conclusions and lays out future work.

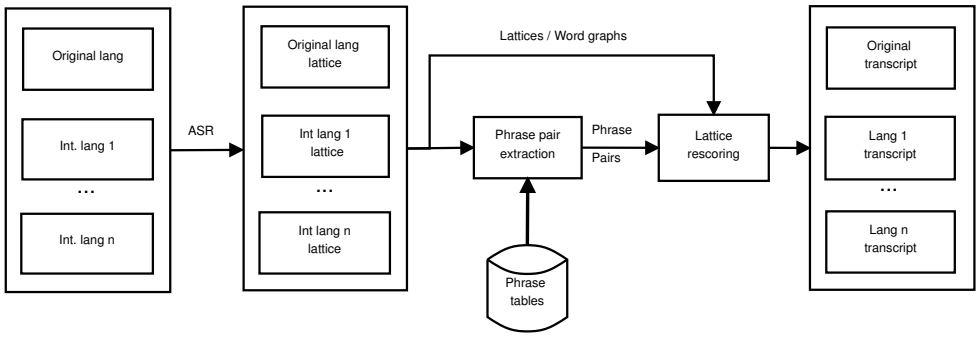

Figure 1: The proposed system architecture

2. Proposed method

2.1. Architecture

The overall system architecture is described in Figure 1. Our proposed method consists of the following steps:

- Collect phrase table pairs for each of the language pairs that we wish to combine.
- Using an automatic speech recognizer, transcribe the speech in the original language, as well as its simultaneously intepreted versions, generating the best as well as alternative hypotheses in the form of *lattices*, together with posterior n-gram distributions.
- For each language pair, intersect the lattices with the respective phrase table, obtaining a set of phrase pairs common to both the lattices and the phrase table.
- Score the obtained phrase pairs and select a subset of these that will be used to rescore the lattices and obtain a new transcription.

2.2. ASR and SMT systems description

For our experiments, we considered three languages, English, Portuguese and Spanish, with the source language of the speeches being English. We used Audimus [6], a hybrid ANN-MLP WFST-based recognizer, as the ASR engine. The feature extraction front-end includes MSG, PLP and RASTA streams. We trained 4-gram language models for each of the languages using the Europarl Parallel Corpus [7], and used our existing acoustic models and lexica for these three languages [8]. We also generated phrase tables for all the possible language combinations (Portuguese-Spanish, Portuguese-English, and Spanish-English), with the Moses toolkit [9], again using the Europarl Parallel Corpus.

2.3. Intersection between lattices and phrase tables

The intersection step consists of finding those phrase pairs $source \mid\mid\mid target$ which simultaneously belong to the phrase table, for which $source$ can be found in the source lattice and $target$ in the target lattice. Additionally, the source and target phrases must be found close together in the respective lattices, where $\delta = 10s$ controls the maximum delay that is expected from the interpreters.

The intersection step is potentially computationally expensive, since phrase tables can have millions of entries and lattices produced for a few minutes of speech can have millions of nodes and edges. Simply enumerating and comparing phrases in the lattices and phrases table would not be feasible, so we developed a number of techniques to alleviate this problem. First, the phrase tables are pruned to eliminate very low probabity entries, reducing their size to about one half of the original. Then, we preprocess the phrase table into a tree, by inserting each of the phrase pairs. To insert a phrase pair into the tree, we first insert each word of the source phrase, starting at the root, and then each word of the target phrase. Then, we process each of the lattice nodes into a tree of depth k, where k is the length of the longest phrase to be considered, by adding to the tree all sequences of k words or less that can be reached from that node. We also annotate each node n of the tree with the posterior probability of the phrase corresponding to the path from the root to n, together with the end time of that phrase. In this way, the intersection process can be recast as simply walking down the phrase table tree, and the lattice trees, simultaneously. On the source side of the phrase table tree we keep only those paths that are also in the source lattice tree, and on the target side those which are on the target lattice tree.

Figure 2 illustrates the preprocessing and intersection for a simple phrase table, containing the phrase pairs 'IMF ||| FMI', 'European Union ||| UE', and 'European Union ||| União Europeia', a source lattice with 'IMF' and 'INF', and a target lattice with 'FMI' and 'FME'. We begin by considering the branches at the root of the

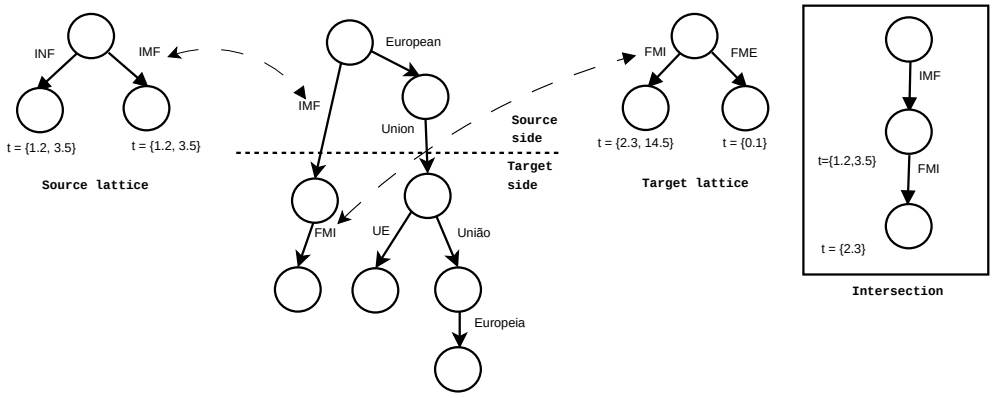

Figure 2: Intersection between an EN-PT phrase table, a EN source lattice, with phrases "IMF" and "INF", and a PT target lattice, with phrases "FMI" and "FME". The phrase table contains the pairs 'IMF ||| FMI', 'European Union ||| UE', and 'European Union ||| União Europeia'. The dashed arrows represent tree branches matched during the intersection process.

source lattice that match with those at the root of the phrase table; only 'IMF' matches, so that is added to the intersection, including the times of occurrence. A leaf of the source lattice has been reached, so the algorithm switches to the root of the target lattice. The only branch that matches with the outgoing branches of the phrase table's current node is the one labeled 'FMI', so that is added to the intersection - note that the entry at time 14.5 is dropped because $\delta > 10$ - and the algorithm terminates. Again, we reached a lattice leaf, and there are no reamining pending decisions to backtrack to.

2.4. Phrase pair selection and rescoring

Clearly, we should not add all the phrase pairs in the calculated intersection, since some of these are likely to have occurred by chance. For instance, a phrase pair such as 'and ||| y' occurs very often in the lattice-phrase table intersection, but not in what has actually been said. Therefore, we train a simple linear model to predict whether a generated phrase pair is actually present in the speech streams, i.e., we compute its score according to the expression $SC(pp) = a_0 + \sum_{i=1}^{N} a_i f_i(pp)$ where the f_i are feature functions, and we select pairs with $SC > 0$.

The feature weights used for tuning the classifier are selected by using Powell's method to optimize WER on the development set. As features, we include:

- posterior probability – the posterior probabilities of each of the phrases in the phrase pair, obtained from the initial recognition pass, are included.
- phrase table features – the phrase pairs are extracted from Moses format phrase tables, which include both direct and inverse translation and lexical probabilities.
- language model scores – the n-gram language model scores of each of the phrases and individual words in the phrases are also included. These help assign a higher weight to phrase pairs that are less likely to have occurred by chance.
- time distances – are based on the fact that closer phrases should be assigned a higher weight.
- number of languages – if several languages agree on the translation for a certain phrase, then it is more likely to be the correct translation. For instance, if we find 'parlamentario' and 'parlamentar' in both Spanish and Portuguese lattices, then that is a stronger indication that the word 'parliamentary' is in the English version than if it only appeared in one of the languages.

The final step consists in an A* rescoring of the lattices produced by the recognizer. The selected phrases are assigned an additive bonus (in the log-score domain) when computing language model scores during search, therefore biasing the decoder towards these phrases. This bonus is only a function of the number of words of the phrase, and is determined together with model weight tuning for the phrase pair selection step. The time stamps derived during the intersection step are used to ensure that we only assign a bonus to occurrences of phrase pairs at the appropriate times.

3. Results

We collected and transcribed two sets of English speeches (both from native and non-native speakers) from the Environment, Public Health and Food Safety (ENVI), Development (DEVE), Internal Market and Consumer Protection (IMCO) and Legal Affairs (LEGAL) committees. The first of these sets was used for tuning the parameters for phrase pair selection mentioned in section 2, whereas the other, consisting of 4 speeches, was used for test-

Speech	EN	+PT	+ES	+PT+ES
DEVE	24.54%	22.33%	21.82%	20.40%
ENVI	20.60%	17.83%	18.84%	16.28%
IMCO	35.12%	31.03%	33.00%	29.97%
LEGAL	33.76%	31.43%	32.45%	28.42%
Average	28.50%	25.65%	26.52%	23.77%

Table 1: WER for the 4 speeches. The 1^{st} column is the error of the baseline system , the 2^{nd} and 3^{rd} the WER of the English original speech after combining with the Portuguese and Spanish interpretations, respectively, and the 4^{th} the error after combining with both interpretations.

Speech	EN	1^{st} iter	2^{nd} iter
DEVE	24.54%	20.40%	18.35%
ENVI	20.60%	16.28%	14.70%
IMCO	35.12%	29.97%	29.79%
LEGAL	33.76%	28.42%	27.38%
Average	28.50%	23.77%	22.56%

Table 2: WER for two iterations of the system (3^{rd} column) compared with the baseline system (1^{st} column) and the first iteration (2^{nd} column)

ing. We also collected the corresponding Spanish and Portuguese interpreted versions of these speeches.

We then executed the proposed method. Table 1 summarizes the main results. We observe improvements in the English WER, relative to the baseline, both when using only one interpreted language and when using two interpreted languages. As expected, using two interpreted languages (English and Portuguese, 16.6% relative improvement) to improve the recognition of the original language outperforms using only one language (Portuguese, 10.0%) or (Spanish, 6.9%). The combined perfomance gain from using two languages appears to be only slightly lower than the cumulative improvements of the individual languages, which suggests our approach may scale to higher numbers of languages.

Next, we performed unsupervised speaker adaptation of our English acoustic model using the output from our method as a reference. Then we re-applied our method and observed the results, summarized in Table 2. We achieved a further 4.2% relative improvement, which translates to 20.8% better than the baseline system. The improvements were, however, more significant in the speeches with lower WER, which points to the use of a confidence measure to select higher quality sections of the automatic transcriptions.

4. Conclusions

In this paper we presented an approach which combines multiple views of a speech, in the form of simultaneously interpreted versions, whenever available, to yield improved recognition results. We obtained a 16.3% relative WER improvement (20.8% after a second iteration) when applying this method to speeches of the European Parliament Committees, using English as the original language and Portuguese and Spanish as the interpreted languages.

In future work, we intend to expand our system with increasing numbers of recognizers of different languages, and analyse whether the reduction in WER demonstrated here is incremental . We also would like to extend this approach to OOV and foreign word detection and recovery, since it seems likely that these problems could benefit from the information that can be obtained from multiple streams. Confidence measure estimation is another area where the use of multiple streams could bring added robustness. Finally, while we focused on multiple speech streams only in this paper, our method is trivially applicable to the case where there is a mixture of both speech and text streams, by converting the text streams into single-hypothesis lattices, where it would likely be more effective due to the absence of recognition errors.

5. Acknowledgements

Support for this research was provided by the Fundação para a Ciência e Tecnologia (Portuguese Foundation for Science and Technology) through the Carnegie Mellon Portugal Program under Grant SFRH / BD / 33767 / 2009.

6. References

[1] Lopez, A., "Statistical Machine Translation", in ACM Computing Surveys 40(3), article 8, 2008.

[2] Khadivi, S.; Ney, H., "Integration of Speech Recognition and Machine Translation in Computer-Assisted Translation," , in IEEE Transactions on Audio, Speech, and Language Processing , vol.16, no.8, pp.1551-1564, 2008

[3] Reddy, A.; Rose, R.C, "Integration of Statistical Models for Dictation of Document Translations in a Machine-Aided Human Translation Task," in IEEE Transactions on Audio, Speech, and Language Processing, vol.18, no.8, pp.2015-2027, 2010

[4] Paulik, M., Stüker, S., Fügen, C., Schultz, T., Schaaf, T. and Waibel, A. , "Speech Translation Enhanced Automatic Speech Recognition", In Proc. of the ASRU, San Juan, Puerto Rico, 2005

[5] Paulik, M., Waibel, A., "Extracting Clues from Human Interpreter Speech for Spoken Language Translation", in Proc. of ICASSP, Las Vegas, USA, 2008.

[6] Meinedo, H., Caseiro, D. A., Neto, J. P., Trancoso, I., AUDIMUS.media: a Broadcast News speech recognition system for the European Portuguese language, in Proc. of PROPOR, Faro, Portugal, 2003.

[7] Koehn, P., "Europarl: A Parallel Corpus for Statistical Machine Translation", in Proc. of MT Summit , 2005 .

[8] Meinedo, H., Abad, A., Pellegrini, T., Trancoso, I., Neto, J. P., "The L2F Broadcast News Speech Recognition System", in Proc. of Fala2010, Vigo, Spain, 2010.

[9] Koehn,P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., Herbst, E., "Moses: Open Source Toolkit for Statistical Machine Translation", in Proc. of the ACL demo session, Prague, Czech Republic, 2007.