SPEECH TRANSLATION ENHANCED AUTOMATIC SPEECH RECOGNITION

M. Paulik[1,2], *S. Stüker*[1], *C. Fügen*[1], *T. Schultz*[2], *T. Schaaf*[2], *and A. Waibel*[1,2]

Interactive Systems Laboratories

[1]Universität Karlsruhe (Germany), [2]Carnegie Mellon University (USA)

{paulik, stueker, fuegen, waibel}@ira.uka.de, {tschaaf, tanja}@cs.cmu.edu

ABSTRACT

Nowadays official documents have to be made available in many languages, like for example in the EU with its 20 official languages. Therefore, the need for effective tools to aid the multitude of human translators in their work becomes easily apparent. An ASR system, enabling the human translator to speak his translation in an unrestricted manner, instead of typing it, constitutes such a tool. In this work we improve the recognition performance of such an ASR system on the target language of the human translator by taking advantage of an either written or spoken source language representation. To do so, machine translation techniques are used to translate between the different languages and then the invovled ASR systems are biased towards the gained knowledge. We present an iterative approach for ASR improvement and outperform our baseline system by a relative word error rate reduction of 35.8% / 29.9% in the case of a written / spoken source language representation. Further, we show how multiple target languages, as for example provided by different simultaneous translators during European Parliament debates, can be incorporated into our system design for an improvement of all involved ASR systems.

1. INTRODUCTION

The recently enlarged European Union has 20 official languages. Official language means that all official EU documents have to be translated into these languages. Therefore, the need for effective tools to aid the multitude of human translators in their work becomes easily apparent. An automatic speech recognition (ASR) system, enabling the human translator to speak his translation in an unrestricted manner, instead of typing it, constitutes such a tool. Dymetman et. al [1] and Brown et. al [2] proposed to improve the recognition performance of such an ASR system in the case of a given source language document. They used machine translation (MT) techniques for improving the target language ASR system for the human translator with the help of the information given in the source language document. Based on this idea, we developed in our previous work [3] an iterative approach for improving the recognition performance of such an ASR system for the human translator. Figure 1(a) depicts the overall iterative system design. As this system relies on the availability of the source documents translated by the human translator, we called our approach document driven machine translation enhanced ASR (MTE-ASR). The key idea of this iterative system design is to recursively apply the improved ASR output to enhance the involved machine translation system for a further ASR improvement.

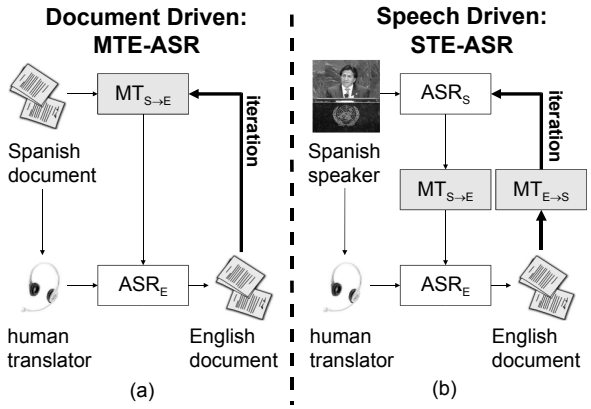

Fig. 1. Machine Translation and Speech Translation Enhanced ASR

In this work we extend our iterative system design to the case where only a spoken representation of the source language is available, as it may be the case for simultaneous translations provided during a European Parliament Plenary Session. Such a speech translation enhanced ASR system (STE-ASR) is shown in Figure 1(b). We will show that the presented iterative speech driven approach is scalable to not just one additional audio stream, but to many audio streams in multiple languages and that it automatically provides an improvement in recognition accuracy of all involved ASR systems. Therefore, it is particularly suited for debates where the speech of a speaker is simultaneously translated

This work has been funded in part by the European Union under the integrated project TC-Star -Technology and Corpora for Speech to Speech Translation - (IST-2002-FP6-506738, http://www.tc-star.org).

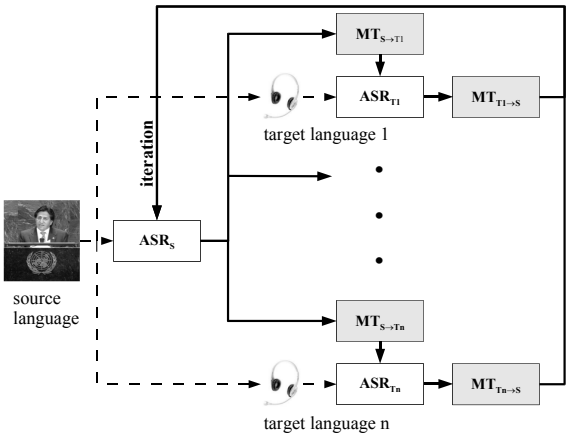

Fig. 2. STE-ASR in the case of n target languages.

into multiple languages. Given one STE-ASR system for each of the simultaneous translators as well as the speaker, it is possible to directly create high quality transcripts of the debate in all used languages, so that only a minimal amount of post-editing of the automatically created transcripts is necessary. Figure 2 shows a scenario in which the multiple audio streams of the human simultaneous translators are used for an improvement of the one source language ASR system.

2. BASELINE

2.1. Data

As before in [3] we are using Spanish as source language and English as target language. The used data set consists of 500 parallel English and Spanish sentences in form and content close to the Basic Travel Expression Corpus (BTEC) [4]. The sentences were presented two times, each time read by three different Spanish and five different English speakers. Ten percent of the data was randomly selected as held-out data for system parameter tuning. Parameter tuning was done by manual gradient descent throughout this work. Because of some flawed recordings, the English data set has 880 sentences with 6,751 (946 different) words. The respective Spanish data set has 900 sentences composed of 6,395 (1,089 different) words. The Spanish audio data equals 45 minutes, the English 33 minutes.
Since the sentences were presented two times there are always two ASR hypotheses for each sentence, decoded on the speech of two different speakers. Using both of these hypotheses within our iterative system design would change the system into a voting system that choses between these two hypotheses. For this reason, the data set was split into two disjoint parts, so that each Spanish-English sentence pair occurs only once within each subset. Based on these two subsets, two different iterative STE-ASR systems had to be examined. In the following only the average performance, calculated on the two individual system results, is given.

	WER	OOV	Perplexity
English Baseline ASR	20.4	0.53%	86.0
Spanish Baseline ASR	17.2	2.04%	130.2

Table 1. Performance characteristics of the baseline ASR systems.

2.2. Baseline ASR Systems

For the ASR experiments in this work we used the Janus Recognition Toolkit (JRTk) featuring the IBIS single pass decoder [5]. Table 1 gives an overview on the performance characteristics of the English and Spanish baseline ASR system.
The English speech recognition system is a sub-phonetically tied semi-continuous three-state HMM based system that has 6K codebooks, 24K distributions and a 42-dimensional feature space on MFCCs after LDA. It uses semi-tied covariance matrices, utterance-based CMS and incremental VTLN with feature-space constrained MLLR. The vocabulary size is 18K. The recognizer was trained on 180h Broadcast News data and 96h Meeting data. The back off tri-gram language model was trained on the English BTEC which consists of 162.2K sentences with 963.5K running words from 13.7K distinct words.
The Spanish recognizer has 2K codebooks and 8K distributions; all other main characteristics are equivalent to the characteristics of the English recognizer. The vocabulary size is 17K. The system was trained on 112h South American speech data (mainly Mexican and Costa Rican dialects) and 14h Castilian speech data. The South American corpus was composed of 70h Broadcast News data, 30h Globalphone data and 12h Spanish Spontaneous Scheduling Task data. The back-off tri-gram LM was trained on the Spanish part of the BTEC.

2.3. Baseline MT Systems

The ISL statistical machine translation system [6] was used for creating the English-to-Spanish and Spanish-to-English translations. This MT system is based on phrase-to-phrase translations (calculated on word-to-word translation probabilities) extracted from a bilingual corpus, in our case the Spanish/English BTEC. It produces an n-best list of translation hypotheses for a given source sentence with the help of its translation model (TM), target language model and translation memory. The translation memory works as follows: for each source sentence that has to be translated the closest matching source sentence, with regard to the edit distance,

is searched in the training corpus and extracted along with its translation. In case of an exact match the extracted translation is used, otherwise different repair strategies are applied to find the correct translation. The translation model computes the phrase translation probability based on word translation probabilities found in its statistical IBM1 forward and backward lexica regardless of the word order. The word order of the MT hypotheses is therefore appointed by the LM and translation memory. Since the MT and the ASR use the same language models, only the translation memory can provide additional word order information for improving the ASR.

3. ASR IMPROVEMENT TECHNIQUES

The ASR improvement techniques applied within our iterative system design are a combination of up to three different basic ASR improvement techniques. A short overview on these three basic ASR improvement techniques is given in this chapter. For a more elaborate description refer to [3].

3.1. Hypothesis Selection by Rescoring

For hypothesis selection the 150 best ASR hypotheses of the ASR system are used together with the first best MT hypothesis of the MT system preceding this ASR system within the iterative cycle. The applied rescoring algorithm computes new scores (negative log-probabilities) for each of the 151 sentences by summing over the weighted and normalized ASR score (s_{ASR}), language model score (s_{LM}), and translation model score (s_{TM}) of this sentence. To compensate for the different ranges of the values for the TM, LM and ASR scores, the individual scores in the n-best lists are scaled to $[0; 1]$.

$$s_{final} = s'_{ASR} + w_{LM} * s_{LM} + w_{TM} * s_{TM} \quad (1)$$

The ASR score output by the JRTk is a linear combination of acoustic score, scaled language model score, word penalty lp and filler word penalty fp. The language model score within this linear combination contains discounts for special words or word classes. The rescoring algorithm allows to directly change the word penalty and the filler word penalty added to the acoustic score. Moreover, four new word context classes with their specific LM discounts are introduced: MT mono-, bi-, trigrams and complete MT sentences (the respective LM discounts are md, bd, td and sd). MT n-grams are n-grams included in the respective MT n-best list; MT sentences are defined in the same manner. The ASR score in equation (1) is therefore computed as:

$$\begin{aligned} s'_{ASR} = & s_{ASR} + lp' * n_{words} + fp' * n_{fillerwords} \\ & - md * n_{MTmonograms} - bd * n_{MTbigrams} \\ & - td * n_{MTtrigrams} - sd * \delta_{isMTsentence} \end{aligned} \quad (2)$$

The rescoring approach applies MT knowledge in two different ways: by computing the TM score for each individual hypothesis and by introducing new word class discounts based on MT n-best lists. Our former experiments conducted in [3] have shown that the MT mono-gram discounts have the strongest influence on the success of the rescoring approach, followed by the TM score. Other parameters apart from the mono-gram discount md and translation model weight w_{TM} only have inferior roles and can be set to zero. This suggests that the additional word context information in form of MT bi- and tri-grams is not very useful for improving the ASR. However, the MT component is very useful as a provider for a "bag-of-words" that predicts which words are going to be used by the human translator.

3.2. Cache Language Model

A classical cache language model has a dynamical memory component that remembers the recent word history of m words to adjust the language model probabilities based on this history. The cache LM used in our system has a dynamically updated 'cache' whereas the LM probabilities are influenced by the content of this cache. However, the cache is not used to remember the recent word history but to hold the words (mono-grams) found in the respective MT n-best list of the sentence that is being decoded at the moment. Our cache LM is realized by defining the members of the word class mono-gram in the same manner as for the rescoring approach, but now dynamically, during decoding. Within the basic ASR improvement techniques, the cache LM approach yields the best improvements results, closely followed by the rescoring approach. This result once again validates the usefulness of the "bag-of-words" knowledge provided by the MT. As this "bag-of-words" knowledge is already applied during decoding, new correct hypotheses are found due to positive pruning effects. This explains why the cache LM approach is able to slightly outperform the rescoring approach, although it lacks the additional form of MT knowledge used by the rescoring approach, namely the direct computation of the TM score.

3.3. Language Model Interpolation

For language model interpolation, the original LM of the ASR system is interpolated with a small back-off tri-gram language model computed on the translations found within all MT n-best lists. LM interpolation yields only small improvements compared to the cache LM and the rescoring approach. This can be explained by the little value of MT word

context information for ASR improvement already stated in 3.2.

4. MT IMPROVEMENT TECHNIQUES

Similar to the improvement of the ASR, the MT improvement technique within our iterative system design is a combination of two basic MT improvement techniques, namely language model interpolation and MT system retraining. For language model interpolation, the original MT language model is interpolated with a small back-off tri-gram language model computed on the hypotheses found within all ASR n-best lists. MT system retraining is done by adding the ASR n-best lists several times to the original training data and computing new IBM1 lexica (forward and backward lexicon), whereas the translation memory component of the MT system is held fixed to the original training data. The reason for keeping the translation memory fixed is that an updated memory leads to a loss of complementary MT knowledge that is valuable for further ASR improvement. An updated memory sees to it that the ASR n-best hypotheses added to the original training data are chosen as translation hypotheses by the MT system, meaning that only a slightly changed ASR output of the preceding iteration is used for ASR improvement in the next iteration instead of new MT hypotheses.
The LM interpolation contributes the most to the MT improvement if the translation memory is kept fix. This means that, while the word context information provided by the MT is of only minimal use for improving the ASR, word context information provided by the ASR is very valuable to improving the MT.

5. DOCUMENT DRIVEN CASE: MTE-ASR

Different combinations of the basic ASR and MT improvement techniques described in section 3 and 4 were taken into consideration for the final document driven system design. The best results in regard to improving the English ASR system were observed when using the combination of LM interpolation and retraining with a fixed translation memory as MT improvement technique. The combination of rescoring and cache LM in iteration 0 and the combination of rescoring, cache LM and interpolated LM in iteration 1 yielded the best results as ASR improvement techniques. The better performance resulting from the additional use of LM interpolation after iteration 0 is due to the improved MT context information. The success of the subsequent rescoring of the ASR output is due to the additional form of MT knowledge applied by the rescoring approach; in contrast to the cache LM approach, rescoring does not only consider the MT "bag-of-words" knowledge but also considers the TM score. In fact, it could be observed that the most important parameter for rescoring on cache LM system output was the translation model weight w_{TM}, since after setting all other parameter to zero, still similar good results could be achieved. No significant improvements were observed for iterations > 1. This was true for all examined system combinations that applied a subsequent rescoring on the ASR system output. If no rescoring was used, similar results to the case where rescoring was used could be obtained, but only after several (> 3) iterations. Figure 3 gives an overview on the components of our final document driven iterative system design along with the respective performance values. With the iterative approach we were able to reduce the WER of the English baseline ASR system from 20.4% to 13.1%. This is equivalent to a relative reduction of 35.8%.

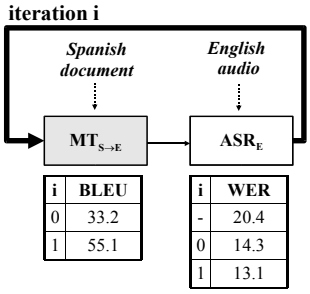

Fig. 3. MTE-ASR; performance of the involved system components in iteration 0 and 1. The performance of the baseline ASR system is marked as iteration '-'.

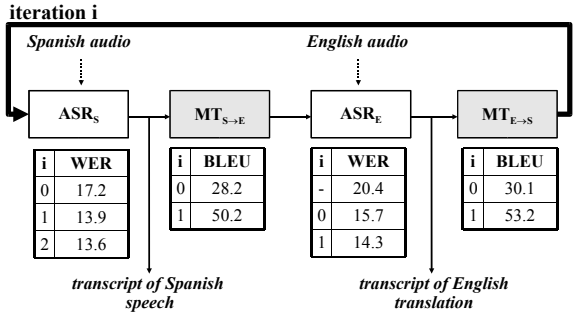

Fig. 4. STE-ASR; performance of the involved system components in iteration 0 and 1. The performance of the English baseline ASR system is marked as iteration '-'.

6. SPEECH DRIVEN CASE: STE-ASR

6.1. Improvement of Target Language Side ASR

Different combinations of the basic ASR and MT improvement techniques were taken into consideration for the final speech driven system design. It turned out that exactly the same combinations as for the document driven case yielded the best results. As in the document driven case,

it was sufficient to improve the MT components just once within the iterative system design for gaining best results in speech recognition accuracy (for both involved ASR systems). This means that in order to avoid overfitting, the iterative process should be aborted right before an involved MT component would be improved a second time. Figure 4 gives an overview of the components of our final speech driven iterative system design along with the respective performance values. The WER of the English baseline ASR system was reduced from 20.4% to 14.3%. This is a relative reduction of 29.9%.

In iteration 0, the BLEU score of the Spanish-to-English MT system is 15.1% relative worse than in the document driven case. This is due to the fact that the Spanish source sentences used for translation now contain speech recognition errors. In this context it should be noted that this loss in MT performance is of approximately the same magnitude as the WER of the Spanish input used for translation, i.e. it is of approximately the same magnitude as the WER of the Spanish baseline system. The loss in MT performance leads to a smaller improvement of the English ASR system compared to the document driven case. However, the loss in MT performance does not lead to a loss in English speech recognition accuracy of the same magnitude; compared to the document driven case the WER of the English ASR system is only 9.8% relative higher. Figure 5 shows a detailed comparison of the performance of the English ASR system in the document driven and the speech driven case. Even though the gain in recognition accuracy is already remarkably high in both cases without applying any iteration, a still significant gain in performance is to be observed in the first iteration.

As already mentioned in section 2.1, we are in fact using two different STE-ASR systems, one for each of the two data subsets. Figure 6 shows the best and worst performing speakers within the two English ASR subsystems before applying MT knowledge and after applying MT knowledge with the help of our iterative scheme. While the WER of the worst speaker is reduced by 36.7% relative, the WER of the best speaker is only reduced by 31.3% relative. This means that for speakers with higher word error rates a higher gain in recognition accuracy is accomplished by applying MT knowledge.

6.2. Improvement of Source Language Side ASR

The ASR driven system design automatically provides an improvement of the involved source language ASR. The WER of the Spanish baseline ASR of 17.2% is reduced by 20.9% relative. This smaller improvement in recognition accuracy compared to the improvement of the English ASR may be explained by the fact that Spanish is a morphological more complicated language than English.

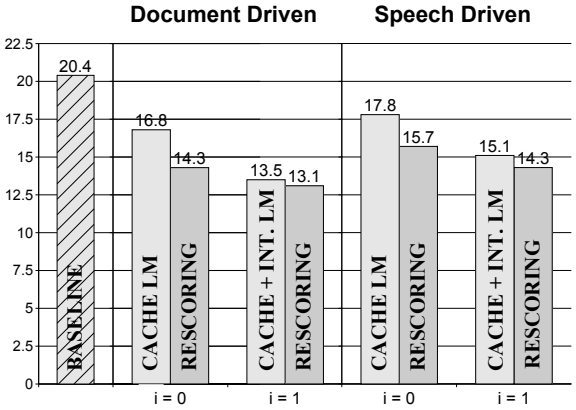

Fig. 5. Detailed comparison of MTE-ASR and STE-ASR.

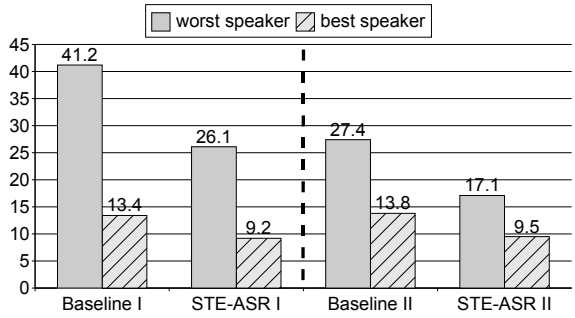

Fig. 6. Development of WERs for different speakers within the two STE-ASR subsystems.

7. MULTIPLE LANGUAGE SOURCES

As already mentioned at the beginning, it is directly possible to incorporate not just one, but several target language audio streams into our iterative system design. For this, the applied improvement techniques only need to be adapted minimally. The adaption of the cache LM approach as well as the LM interpolation (for ASR and MT improvement) and MT retraining is done by including all MT/ASR n-best lists of the preceding MT/ASR systems in the iterative cycle. For rescoring, Equation 1 is extended to allow for several TM scores provided by several MT systems with different target languages, i.e. instead of one TM score and associated TM weight we have now up to n TM scores with their respective TM weights. In the following, we show how an already speech translation enhanced English ASR system is further improved by adding knowledge provided by one additional audio stream in a different target language.

7.1. Baseline

For this set of experiments we used a BTEC held-out data set consisting of 506 parallel Spanish, English and Mandarin Chinese sentences. Ten percent of the data was randomly selected for system parameter tuning. The English

	WER	OOV	Perplexity
English Baseline ASR	13.5	0.56%	21.9
Spanish Baseline ASR	15.1	3.20%	75.5
Mandarin Baseline ASR	20.0	1.14%	70.1

Table 2. Performance characteristics of the baseline ASR systems on the BTEC held-out data set.

and Spanish sentences were read twice, the Chinese Sentences were read just once. The same Spanish and English baseline ASR systems were used as before. For Chinese speech recognition we used the ISL RT04 Mandarin Broadcast News evaluation system [7]. The vocabulary of the Chinese ASR system has 17K words. The Chinese LM was computed on the Chinese BTEC. Table 2 gives an overview of the performance of the baseline ASR systems.

7.2. STE-ASR Results

Initially we used only the Spanish and English audio streams for speech translation based ASR improvement. We applied the same iterative STE-ASR technique as in section 6 with the exception that no LM interpolation was used for improving the English ASR system, as a slightly worse WER was observed for doing so. The negative influence of LM interpolation on the performance of the English ASR system can be explained by the already very good match of the English baseline LM with the used data set (the perplexity is only 21.9). The WER of the Spanish ASR system was reduced from 15.1% to 13.4%. The WER of the English ASR system was reduced from 13.5% to 10.6%. Next, we examined if the performance of the improved English ASR system can be further increased by taking advantage of the additional Chinese audio stream. For this, we first improved the Chinese baseline system with the help of the latest computed English system output and we then used the output of the improved Chinese system to once again improve the English system. The MT systems for translating between English and Chinese were trained on the Chinese-English BTEC. The accomplished BLEU scores were with 21.2 for E $\rightarrow$ C and with 24.1 for C $\rightarrow$ E very moderate. Nevertheless, we were able to reduce the WER of the Chinese system from 20.0% to 17.1% and for the English system from 10.6% to 10.3%. Although statistically insignificant, the reduction for the English system constitutes a very promising result in the context of multiple target language STE-ASR.

8. SUMMARY

In this work we successfully extended our iterative approach for ASR improvement in the context of human-mediated translation scenarios to the case where only spoken language representations are available. One key feature of our iterative STE-ASR design is, that the recognition accuracy of all involved ASR system is automatically improved, i.e. not only the target language ASR but also the source language ASR is improved. Using Spanish as source language and English as target language, we were able to reduce the WER of our English baseline ASR system by 29.9% relative and the WER of our Spanish baseline system by 20.9%. Further, we showed that the extension of our former document driven MTE-ASR approach to the speech driven case enables us to directly incorporate not just one, but multiple target language audio streams, as they may be available for example from several simultaneous translators during a United Nations or European Parliament session. Our future work will focus on the incorporation of one or more additional target language audio streams as well as the the adaption of our current system to a more realistic data set, like for example European Parliament Plenary Sessions data.

9. REFERENCES

[1] M. Dymetman, J. Brousseaux, G. Foster, P. Isabelle, Y. Normandin, and P. Plamondon, "Towards an automatic dictation system for translators: the transtalk project," in *Proceedings of ICSLP*, Yokohama, Japan, 1994.

[2] P. Brown, S. Della Pietra S. Chen, V. Della Pietra, S. Kehler, and R. Mercer, "Automatic speech recognition in machine aided translation," in *Computer Speech and Language*, 8, 1994.

[3] M. Paulik, C. Fügen, S. Stüker, T. Schultz, T. Schaaf, and A. Waibel, "Document driven machine translation enhanced asr," in *Proceedings of the 9th European Conference on Speech Communication and Technology*, Lisbon, Portugal, September 2005.

[4] G. Kikui, E. Sumita, T. Takezawa, and S. Yamamoto, "Creating corpora for speech-to-speech translation," in *Proceedings of Eurospeech*, Geneve, Switzerland, 2003.

[5] H. Soltau, F. Metze, C. Fügen, and A. Waibel, "A one-pass decoder based on polymorphic linguistic context assignment," in *Proceedings of ASRU*, Madonna di Campiglio, Italy, 2001.

[6] S. Vogel, S. Hewavitharana, M. Kolss, and A. Waibel, "The isl statistical machine translation system for spoken language translation," in *Proceedings of IWSLT*, Kyoto, Japan, 2004.

[7] H. Yua, Y. Tam, T. Schaaf, S. Stüker, Q. Jin, M. Noamany, and T. Schultz, "The isl rt04 mandarin broadcast news evaluation system," in *EARS Rich Transcription Workshop*, Palisades, NY, USA, 2004.