Modern Applied Science; Vol. 11, No. 4; 2017
ISSN 1913-1844 E-ISSN 1913-1852
Published by Canadian Center of Science and Education

An Investigation into Methodology and Metrics Employed to Evaluate the (Speech-to-Speech) Way in Translation Systems

Parnyan Bahrami Dashtaki[1]

[1] Literature in English, Kakatiya University

Correspondence: Parnyan Bahrami Dashtaki, Literature in English, Kakatiya University, E-mail: bahrami949494@gmail.com

Received: February 2, 2016 Accepted: May 2, 2016 Online Published: February 8, 2016

doi:10.5539/mas.v11n4p55 URL: http://dx.doi.org/10.5539/mas.v11n4p55

Abstract

Speech-to-speech translation is a challenging problem, due to poor sentence planning typically associated with spontaneous speech, as well as errors caused by automatic speech recognition. Based upon a statistically trained speech translation system, in this study, we try to investigate methodologies and metrics employed to assess the (speech-to-speech) way in translation systems. The speech translation is performed incrementally based on generation of partial hypotheses from speech recognition. Speech-input translation can be properly approached as a pattern recognition problem by means of statistical alignment models and stochastic finite-state transducers. Under this general framework, some specific models are presented. One of the features of such models is their capability of automatically learning from training examples. The speech translation system consists of three modules: automatic speech recognition, machine translation and text to speech synthesis. Many procedures for incorporation of speech recognition and machine translation have been projected. In this research, we want explore methodologies and metrics employed to assess the (speech-to-speech) way in translation systems.

Keyword: Methodology, speech to speech, translation systems

1. Introduction

A Speech-to-Speech Translation (SST) system is composed of an Automatic Speech Recognizer (ASR) chained to a Spoken Language Translation (SLT) module and to a Text-To-Speech (TTS) component in order to produce the speech in the target language (Hamon & Mostefa, 2008). Speech-to-speech translation is a challenging problem, due to poor sentence planning typically associated with spontaneous speech, as well as errors caused by automatic speech recognition. Most speech translation systems reported in the literature operate within more or less restricted domains (Levin et al., 2000; Frederking et al., 2002; Gao et al., 2002; Rayner and Bouillon, 2002). Many are based on the Interlingua approach to translation; however, systems differ in their linguistic complexity. Knowledge-lean statistical machine translation approaches are nearly universally embraced for the task of unrestricted text translation (Koehn et al., 2003), perhaps because it is more difficult to effectively exploit knowledge in the broad domain. In restricted domains, rule-based and statistical-based approaches clearly show different strengths and weaknesses, which make them complement each other nicely (Wang & Seneff, 2004).

Moreover, the translation module of a speech translation system, a natural off-spring of text-input based translation system, usually takes a single-best recognition hypothesis transcribed in text and performs standard text-based translation. Lots of supplementary information available from speech recognition, such as N-best recognition hypotheses, likelihoods of acoustic and language models, is not well utilized in the translation process. The information can be effective for improving translation quality if employed properly. The supplementary information can be exploited by a tight coupling of speech recognition and machine translation (Ney, 1999) or keeping the cascaded structure unchanged but using an integration model, log-linear model, to re-score the translation hypotheses (Zang et al., 2004).

1.1 Speech Translation

The goal of the speech translation system research is to make straightforward real-time, interpersonal communication via usual spoken language for people who do not share a neutral language. Speech Translation (ST) is the process which spoken expressions are rapidly translated and spoken clearly in a second language. This is in contrast from phrase translation method, where the system merely translates a predetermined and finite

set of sentences that have been manually entered into the system. Speech to speech translation technology supports speakers of various languages to interconnect. Thus it provides fabulous value for humankind in terms of science, cross-cultural interchange and global business. Nowadays, speech translation systems are used all through the domain. Examples include medical facilities, police, schools, retail stores, hotels, and factories. These systems are applicable everywhere that spoken language is being used to communicate (Sangeetha & Jothilakshmi, 2015).

Speech translation is in many respects a particularly difficult version of the translation task. High quality output is essential: the speech produced must sound natural if it is to be easily comprehensible. The quality of the translation itself must also be high, in spite of the fact that, by the nature of the problem, no post-editing is possible. Things are equally difficult on the input side: pre-editing, too, is difficult or impossible, yet ill-formed input and recognition errors are both likely to be quite common. Thus robust analysis and translation are also required. Furthermore, any attempted solutions to these problems must be capable of operating at a speed close enough to real time that users are not faced with unacceptable delays (Carter et al, 1997)

Speech translation (ST) is an important technology for cross lingual (one-way or two-way) oral communication, whose societal role is rapidly increasing in the modern global and interconnected informational age. ST technology as key enabler of universal translation is one of the most promising and challenging future needs and wants in the coming decade (Treichler, 2009).

A ST system consists of two major components: automatic speech recognition (ASR) and machine translation (MT). Over the past years, significant progress has been made in the integration of these two components in the end-to-end ST task (Casacuberta et al, 2008; Matsoukas et al, 2007; Matusov et al, 2006; Ney, 1999; Wang et al, 2010; Zhou et al, 2007).

In the study of Ney (1999), a Bayes-rule-based integration of ASR and MT was proposed, in which the ASR output is treated as a hidden variable. In the study of Zang et al (2004) a log-linear model was proposed to directly model the posterior probability of the translated output given the input speech signal, where the feature functions are derived from the overall outputs of the ASR model, the translation model, and the Part-of-Speech language model. This set of work is later extended with the use of the phrase-based MT component and a lattice/confusion-network based interface between ASR and MT (Matusov, et al, 2005; Saleem et al, 2004).

1.1.1 Speech Translation Systems

A general framework for ST is illustrated in Fig. 1. The input speech signal X is first fed into the ASR module. Then the ASR module generates the recognition output set $\{F\}$, which is in the source language. The recognition hypothesis set $\{F\}$ is finally passed to the MT module to obtain the translation sentence E in the target language (He et al, 2011)

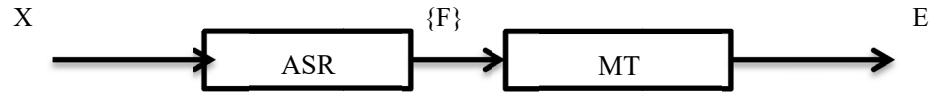

Figure 1. Two components of a full speech translation system

Currently, speech translation technology is available as a product that instantly interprets free form multi-lingual conversations. These systems instantly convert uninterrupted speech. Challenges in achieving this task include overcoming speaker dependent changes in fashion of speaking or pronunciation are issues that have to be dealt with so as to give a high quality translation for every user. Moreover, automatic speech recognition systems have to be prepared to tackle external factors such as acoustic noise or speech by other speaker in real-world use of speech translation systems. For the motivation that the client does not know the target language when speech translation is used, a method need be delivered to the user to check whether the translation is correct, by such means as translating it again back into the user's language (Satoshi, 2009).When we consider speech-to-speech (S2S) translation systems, several abstract models are possible.

1.1.2 Statistical Framework to Speech Translation

Let x^T_1 be the acoustic representation of an input sentence. The translation of x^T_1 into another language can be formulated as the problem of searching for a sequence of words $\hat{t}_1^{\hat{I}}$ in the target language that maximizes.

$$\hat{t}_1^{\hat{I}} = \arg\max \Pr(t_1^I \mid x_1^T)$$

The maximization is carried out over all possible sequences, t^I_1, of all possible lengths, I. for simplicity purposes; we do not present explicitly the maximization over I. Speech translation can be seen as a two steps process.

$$x_1^T \rightarrow S_1^j \rightarrow t_1^I$$

Where:

S^J_1 is possible decoding of x^T_1 in the source language which can be translated into a sequence of words, t^J_1 in the target language. Consequently (Casacuberta et al, 2004)

$$\arg\max_{t_1^I} \Pr(t_1^I \mid x_1^T) = \arg\max_{t_1^I} \sum_{s_1^I} \Pr(t_1^I, s_1^I \mid x_1^T)$$

1.1.3 Automatic Speech Recognition

Speech Recognition (is also known as Automatic Speech Recognition (ASR), or computer speech recognition) is the process of converting a speech signal to a sequence of words, by means of an algorithm implemented as a computer program (Anusuya & Katti, 2010)

1.1.4 Basic Model of Speech Recognition

Research in speech processing and communication for the most part, was motivated by people s desire to build mechanical models to emulate human verbal communication capabilities. Speech is the most natural form of human communication and speech processing has been one of the most exciting areas of the signal processing. Speech recognition technology has made it possible for computer to follow human voice commands and understand human languages. The main goal of speech recognition area is to develop techniques and systems for speech input to machine. Speech is the primary means of communication between humans. For reasons ranging from technological curiosity about the mechanisms for mechanical realization of human speech capabilities to desire to automate simple tasks which necessitates human machine interactions and research in automatic speech recognition by machines has attracted a great deal of attention for sixty years (Tran, 2000)

Based on major advances in statistical modeling of speech, automatic speech recognition systems today find widespread application in tasks that require human machine interface, such as automatic call processing in telephone networks, and query based information systems that provide updated travel information, stock price quotations, weather reports, Data entry, voice dictation, access to information: travel, banking, Commands, Avoinics, Automobile portal, speech transcription, Handicapped people (blind people) supermarket, railway reservations etc. Speech recognition technology was increasingly used within telephone networks to automate as well as to enhance the operator services. This report reviews major highlights during the last six decades in the research and development of automatic speech recognition, so as to provide a technological perspective (Anusuya & Katti, 2010).

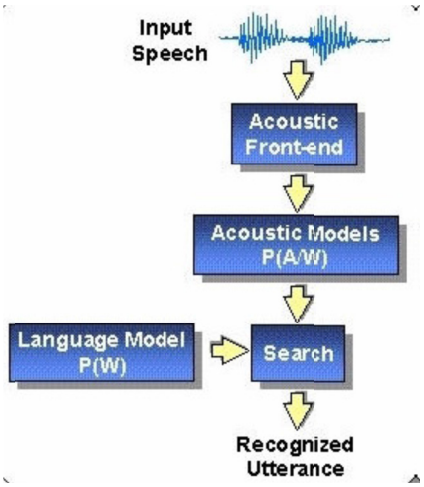

Figure 2. Basic model of speech recognition

Fig.2 shows a mathematical representation of speech recognition system in simple equations which contain front end unit, model unit, language model unit, and search unit. The recognition process is shown below.

1.2 Metrics of Speech to Speech Translation System

Regardless of how many levels and components there are in a given implementation, different metrics could be applied around any input-output pair of interest to help drive quality improvements. In figure 3, for example, we could have a metric around each processing module; that is, one metric for the mapping SLU to SLT, another metric for SLT to TLT, and a third from TLT to TLU. Each metric would be used to study the effectiveness of the system module of interest.

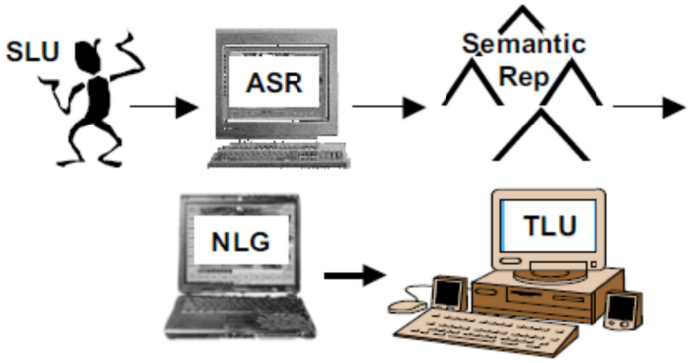

Figure 3. Mmapping SLU to SLT, SLT to TLT and TLT to TLU

In production environments—as opposed to system development—translation metrics are typically applied to random samples of source and target language translations. Metrics based on static reference translations for the automatic evaluation of system quality during system development [Doddington] are thus not in the domain of this discussion. Evaluation is usually performed by a qualified translator with domain knowledge, who is generally employed by a translation agency, though client companies sometimes perform their own internal evaluations (Godden, 2002).

1.3 The BLEU Measure

A fundamental problem of translation evaluation is that there are many possible translations from a source language input to a target language output. The IBM researchers who developed BLEU in 2001 provided a partial solution to this problem by creating test sets with more than one translation for each input. The machine translation output is then compared to these reference translations, and a score is computed based on the number of n-grams in the output that match the references (Condon et al, 2009)

As automated measures are used more extensively, researchers learn more about their strengths and shortcomings, which allows the scores to be interpreted with greater understanding and confidence. Some of the limitations that have been identified for BLEU are very general, such as the fact observed earlier that the measure primarily reflects the accuracy of the words that the system produced with only a brevity penalty to assess what the system may have missed. This makes the measure more like a document similarity measure (Owczarzak, 2007)

1.4 The METEOR Measure

METEOR also addresses another problem that has been associated with BLEU. The ability of BLEU to take into account many possible translations for a given segment of language depends solely on the number of reference translations that are available for comparison. In contrast, METEOR accepts synonyms defined in a resource called Word Net, allowing additional options that are not present in reference translations. For example, METEOR would recognize the equivalence of pain and ache in Figure 4. METEOR also uses stemming to remove inflectional affixes that may prevent translations from matching due to minor variation. For example, after stemming, METEOR would match cries and crying in Figure 1 because they are both forms of the verb cry. However, these enhancements are available only for English: there is no equivalent of WordNet for Iraqi Arabic, and Arabic affixes are often ambiguous out of context, making it difficult to stem words accurately (Condon et al, 2009)

Ref 1: he has some pain in his stomach and always cries and complains about stomach pain
Ref 2: he has some pain in his stomach and he always cries and says I have a stomach pain
Ref 3: he has some stomach pain and always cries saying my stomach hurts
Ref 4: he has a stomach ache and he always cries and says my stomach hurts
System: he has stomach pain and always crying he says pain in stomach

Ref 1: he has some pain in his stomach and always cries and complains about stomach pain
Ref 2: he has some pain in his stomach and he always cries and says I have a stomach pain
Ref 3: he has some stomach pain and always cries saying my stomach hurts
Ref 4: he has a stomach ache and he always cries and says my stomach hurts
System: he has stomach pain and always crying he says pain in stomach

Figure 4. Sample Reference Translations and System Output

The METEOR score is computed by aligning the system output to the closest reference translation as in Figure 6. After stemming, cries and crying are considered a match, as are saying and says. In Figure 5, three words of the reference translation (in boldface) are not matched to the system output, and three words of the system output (not boldface) do not match the reference translation (Condon et al, 2009).

Ref 3: he has some stomach pain and always cries saying my stomach hurts
System: he has stomach pain and always crying he says pain in stomach

Figure 5. METEOR Alignment of System Output and Reference Translation

1.5 The TER, STER and HTER Measures

The TRANSTAC program has also experimented with the TER metric to measure translation quality. Unlike METEOR, TER allows any number of contiguous words to shift positions in a single move. Computation of the TER score is based on the Levenshtein edit distance measure for string matching (Cohen et al, 2003), which counts the number of insertions, deletions, and substitutions required to transform one string into another. Figure 6 shows how the alignment in Figure 6 would be edited to transform the system output into the reference translation. The deletions and substitutions that transform he says pain in into saying could have been aligned differently with no effect on the number of deletions and substitutions (Condon et al, 2009).

Ref 3: he has some stomach pain and always cries saying my stomach hurts
System: he has stomach pain and always crying he says pain in stomach
Edits: insertion substitution deletion substitution deletion substitution deletion

Figure 6. TER Alignment of System Output with Reference Translation and Edits

2. Methodology Used for Automatic Speech Recognition

2.1 IBM's MASTOR

The IBM MASTOR shorthand for Multilingual Automatic Speech-to-Speech Translator is developed for the DARPA CAST and its mission is to develop technologies that facilitate rapid deployment of real-time

Speech-to-Speech Translation of low-resource languages on mobile devices (Gao, et al, 2006).The general structure of MASTOR system has the components of ASR, MT and TTS. This pipelining approach allows system for the deployment of the existing speech and language handing out techniques, while taking care of unique problems in Speech-to-Speech Translation (Dureja and Gautam, 2015)

Grapheme based acoustic models are used to overcome the problem of absence of short vowels Grapheme based acoustic model lead to unambiguous pronunciation of lexicons and hence facilitates the model training and decoding. Also, depending on its context the same grapheme may yield different phonetic sound and lead to less accurate acoustic models. For this reason two different approaches come into existence. The first one is to use short vowels known as full phonetic approach and the second one uses the context-sensitive graphemes in which two different phonemes are generated for the letter "A" (Alif) depending on its position in the word. The IBM ViaVoice product engine is a highly robust and efficient framework which is used for acoustic modelling by using rank based acoustic scores that are derived from tree-clustered context reliant Gaussian Models for both the desktop systems and hand-held systems (Narayanan et al, 2006)

2.2 Verbmobil

Verbmobil is a two way Speech-to-Speech Translation system which does not depend on the speaker. It is used for translation of spontaneous dialogs in mobile situations. It firstly identifies the input and further analyses and translates it, and finally delivers the final translation. This is a multilingual system which handles dialogs delivery in three-business-oriented domains where the translation depends on the context between three languages (German, English and Japanese) (Wahlster, 2013)

This system deals with the spontaneous dialogs. In this case it doesn't mean just continuous speech like in the current dictation systems, but here rational disfluencies and repairing phenomena such as changing mid word, ums and arr, and some short words that are accidently left out in rapid speech are also included in the speech. For example, Verb Mobil corpus has the chance that 20% of all dialog turns having at least one auto-correction and 3% also include false starts. A combined approach for deep and shallow analysis methods is used by this system to find out the slips in the speech and then translate it in accordance to what the person tried to say rather than what was actually said by him (Dureja and Gautam, 2015)

3. Literature Review

Prior work on S2S translation has primarily focused on providing either one-way or two-way translation on a single device (Waibel et al., 2003; Zhou et al., 2003). Typically, the user interface requires the participant(s) to choose the source and target language apriori. The nature of communication, either single user talking or turn taking between two users can result in a one-way or cross-lingual dialog interaction. In most systems, the necessity to choose the directionality of translation for each turn does take away from a natural dialog flow. Furthermore, single interface based S2S translation (embedded or cloud based) is not suitable for cross-lingual communication when participants are geographically distant, a scenario more likely in a global setting. In such a scenario, it is imperative to provide real-time and low latency communication (Bangalore et al, 2012)

Researchers have recognized that translation quality is multi-faceted and that human judgments of even more specific qualities such as fluency and fidelity are not always reliable (King, 1996; Turian, Shen & Melamed, 2003). Given the unevenness and cost of human judgments, researchers have welcomed automated measures such as BLEU and have proposed a plethora of alternative methods, all of which involve comparisons to one or more reference translations (Candon et al, 2008)

In contrast, evaluations of speech translation have relied on human judgments such as the binary or ternary classifications adopted by CMU (Gates et al., 1996) and Verb Mobil (Nübel, 1997) researchers, which combine assessments of accuracy and fluency. Other methods use abstract semantic representations of the source utterances and require human judges to score structural elements of those representations separately. CMU researchers use the Interlingua Interchange Format to represent utterance intent and content (Levin et al., 2000).

Sageetha and Jothilakshmi (2015) conducted a research named "Integrating Machine Translation and Speech Synthesis Component for English to Dravidian Language Speech to Speech Translation System". This paper provides an interface between the machine translation and speech synthesis system for converting English speech to Tamil text in English to Tamil speech to speech translation system. The speech translation system consists of three modules: automatic speech recognition, machine translation and text to speech synthesis. Many procedures for incorporation of speech recognition and machine translation have been projected. Still speech synthesis system has not yet been measured. In this paper, we focus on integration of machine translation and speech synthesis, and report a subjective evaluation to investigate the impact of speech synthesis, machine

translation and the integration of machine translation and speech synthesis components. Here they implement a hybrid machine translation (combination of rule based and statistical machine translation) and concatenative syllable based speech synthesis technique. In order to retain the naturalness and intelligibility of synthesized speech Auto Associative Neural Network (AANN) prosody prediction is used in this work. The results of this system investigation demonstrate that the naturalness and intelligibility of the synthesized speech are strongly influenced by the fluency and correctness of the translated text.

Sanders et al, (2013) conducted a research named "Evaluation methodology and metrics employed to assess the TRANSTAC two-way, speech-to-speech translation systems". One of the most difficult challenges that military personnel face when operating in foreign countries is clear and successful communication with the local population. To address this issue, the Defense Advanced Research Projects Agency (DARPA) is funding academic institutions and industrial organizations through the Spoken Language Communication and Translation System for Tactical Use (TRANSTAC) program to develop practical machine translation systems. The goal of the TRANSTAC program is to demonstrate capabilities to rapidly develop and field free-form, two-way, speech-to-speech translation systems that enable speakers of different languages to communicate with one another in real-world tactical situations without an interpreter. Evaluations of these technologies are a significant part of the program and DARPA has asked the National Institute of Standards and Technology (NIST) to lead this effort. This article presents the experimental design of the TRANSTAC evaluations and the metrics, both quantitative and qualitative, that were used to comprehensively assess the systems' performance.

Brian et al (2011) conducted a research named "Performance Assessments of Two-Way, Free-Form, Speech-to-Speech Translation Systems for Tactical Use". A critical challenge for military personnel when operating in foreign countries is effective communication with the local population. To address this issue, the Defense Advanced Research Projects Agency (DARPA) created the Spoken Language Communication and Translation Systems for Tactical Use (TRANSTAC) program. The program's goal is to develop speech-to speech translation technologies enabling English speakers to quickly communicate with the local population without an interpreter. DARPA has funded the National Institutes of Standards and Technology to lead the design and implementation of the TRANSTAC performance evaluations. This article presents these evaluations that enabled the collection of rich quantitative and qualitative metrics.

He et al, (2011) conducted a research named "WHY WORD ERROR RATE IS NOT A GOOD METRIC FOR SPEECH RECOGNIZER TRAINING FOR THE SPEECH TRANSLATION TASK?" Speech translation (ST) is an enabling technology for cross-lingual oral communication. A ST system consists of two major components: an automatic speech recognizer (ASR) and a machine translator (MT). Nowadays, most ASR systems are trained and tuned by minimizing word error rate (WER). However, WER counts word errors at the surface level. It does not consider the contextual and syntactic roles of a word, which are often critical for MT. In the end-to-end ST scenarios, whether WER is a good metric for the ASR component of the full ST system is an open issue and lacks systematic studies. In this paper, they report recent investigation on this issue, focusing on the interactions of ASR and MT in a ST system. They show that BLEU-oriented global optimization of ASR system parameters improves the translation quality by an absolute 1.5% BLEU score, while sacrificing WER over the conventional, WER-optimized ASR system. They also conducted an in-depth study on the impact of ASR errors on the final ST output. Our findings suggest that the speech recognizer component of the full ST system should be optimized by translation metrics instead of the traditional WER.

Bangalore et al. (2012) conducted a research named "Real-time Incremental Speech-to-Speech Translation of Dialogs". In this work, they addressed the problem of incremental speech-to-speech translation (S2S) that enables cross-lingual communication between two remote participants over a telephone. They investigated the problem in a novel real-time Session Initiation Protocol (SIP) based S2S framework. The speech translation is performed incrementally based on generation of partial hypotheses from speech recognition. They describe the statistical models comprising the S2S system and the SIP architecture for enabling real-time two-way cross-lingual dialog. They presented dialog experiments performed in this framework and study the tradeoff in accuracy versus latency in incremental speech translation. Experimental results demonstrate that high quality translations can be generated with the incremental approach with approximately half the latency associated with non-incremental approach.

Hamon and Mostefa (2008) conducted a research named "An Experimental Methodology for an End-to-End Evaluation in Speech-to-Speech Translation". This paper describes the evaluation methodology used to evaluate the TC-STAR speech-to-speech translation (SST) system and their results from the third year of the project. It follows the results presented in (Hamon et al., 2007), dealing with the first end-to-end evaluation of the project. In this paper, we try to experiment with the methodology and the protocol during the second end-to-end

evaluation, by comparing outputs from the TC-STAR system with interpreters from the European parliament. For this purpose, we test different criteria of evaluation and type of questions within a comprehension test. The results reveal that interpreters do not translate all the information (as opposed to the automatic system), but the quality of SST is still far from that of human translation. The experimental comprehension test used provides new information to study the quality of automatic systems, but without settling the issue of what protocol is best. This depends on what the evaluator wants to know about the SST: either to have a subjective end-user evaluation or a more objective one.

Gao et al. (2006) conducted a research named "IBM MASTOR SYSTEM: Multilingual Automatic Speech-to-speech Translator". In this paper, they described the IBM MASTOR, a speech-to-speech translation system that can translate spontaneous free-form speech in real-time on both laptop and hand-held PDAs. Challenges include speech recognition and machine translation in adverse environments, lack of training data and linguistic resources for under-studied languages, and the need to rapidly develop capabilities for new languages. Another challenge is designing algorithms and building models in a scalable manner to perform well even on memory and CPU deficient hand-held computers. They described their approaches, experience, and success in building working free-form S2S systems that can handle two language pairs (including a low-resource language).

Narayanan et al. (2006) conducted a research named "SPEECH RECOGNITION ENGINEERING ISSUES IN SPEECH TO SPEECH TRANSLATION SYSTEM DESIGN FOR LOW RESOURCE LANGUAGES AND DOMAINS". Engineering automatic speech recognition (ASR) for speech to speech (S2S) translation systems, especially targeting languages and domains that do not have readily available spoken language resources, is immensely challenging due to a number of reasons. In addition to contending with the conventional data-hungry speech acoustic and language modeling needs, these designs have to accommodate varying requirements imposed by the domain needs and characteristics, target device and usage modality (such as phrase-based, or spontaneous free form interactions, with or without visual feedback) and huge spoken language variability arising due to socio-linguistic and cultural differences of the users. This paper, using case studies of rating speech translation systems between English and languages such as Pashto and Farsi, describes some of the practical issues and the solutions that were developed for multilingual ASR development. These include novel acoustic and language modeling strategies such as language adaptive recognition, active-learning based language modeling, class-based language models that can better exploit resource poor language data, efficient search strategies, including N-best and confidence generation to aid multiple hypotheses translation, use of dialog information and clever interface choices to facilitate ASR, and audio interface design for meeting both usability and robustness requirements.

Godden (2002) conducted a research named "Towards a Speech-to-Speech Machine Translation Quality Metric". General characteristics of a pragmatic metric for the *production* evaluation of speech-to-speech translations are discussed. While these characteristics constrain the space of allowable metrics, infinite definition space remains from which to select and define any particular metric. The recommended characteristics are drawn from the author's experience as primary developer of a text-based translation quality metric used in a production environment. The primary contribution is that of strict category ordering and two meta-rules that reduce the variance in assignment of errors to categories.

4. Conclusion

In this paper we investigated the methodology and metrics employed to assess the (speech-to-speech) way in translation systems. We talked briefly about speech translation. Then we introduced speech translation system and the components of it. We described Metrics of speech to speech translation system involved The BLEU Measure, The METEOR Measure and The TER, STER and HTER Measures. We explored Methodology used for automatic speech recognition involved IBM's MASTOR and VERBMOBIL. In the experiments we presented, some methods were applied to translating automatic speech recognition output for English utterances. Based on the Goddon study (2002), U2U (utterance-to-utterance) metric does not automatically become a good metric. The category definitions are of extreme importance, as are the examples used to illustrate the definitions and the training materials created for evaluators. Without clear, unambiguous and precise error definitions no metric will be of any practical value. Hamon and Mostefa (2008) found that that interpreters do not translate all the information (as opposed to the automatic system), but the quality of SST is still far from that of human translation. Bangalore et al demonstrated that high quality translations can be generated with the incremental approach with approximately half the latency associated with nonincremental approach. He et al, (2011) concluded that BLEU-oriented global optimization of ASR system parameters improves the translation quality by an absolute 1.5% BLEU score, while sacrificing WER (word error rater) over the conventional,

WER-optimized ASR system. Sageetha and Jothilakshmi (2015) implemented a hybrid machine translation (combination of rule based and statistical machine translation) and concatenative syllable based speech synthesis technique. The results of this system investigation demonstrate that the naturalness and intelligibility of the synthesized speech are strongly influenced by the fluency and correctness of the translated text.

References

Anusuya, M. A., & Katti, S. K. (2010). Speech recognition by machine, a review. *arXiv preprint arXiv:1001.2267*.

Bangalore, S., Rangarajan Sridhar, V. K., Kolan, P., Golipour, L., & Jimenez, A. (2012,). Real-time incremental speech-to-speech translation of dialogs. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 437-445). Association for Computational Linguistics.

Carter, D., Becket, R., Rayner, M., MacDermid, C., Wiren, M., & Philp, C. (1997). Translation methodology in the spoken language translator: an evaluation. *arXiv preprint cmp-lg/9705015*.

Casacuberta, F., Ney, H., Och, F. J., Vidal, E., Vilar, J. M., Barrachina, S., ... & Nevado, F. (2004). Some approaches to statistical and finite-state speech-to-speech translation. *Computer Speech & Language*, *18*(1), 25-47. https://doi.org/10.1016/s0885-2308(03)00028-7

Cohen, W., Ravikumar, P., & Fienberg, S. (2003, August). A comparison of string metrics for matching names and records. In *Kdd Workshop on Data Cleaning and Object Consolidation*, *3*, 73-78.

Condon, S. L., Phillips, J., Doran, C., Aberdeen, J. S., Parvaz, D., Oshika, B. T... & Schlenoff, C. (2008, May). Applying Automated Metrics to Speech Translation Dialogs. In LREC. https://doi.org/10.1145/1865909.1865959

Dureja, M., & Gautam, S. (2015). Speech-to-Speech Translation: A Review. *International Journal of Computer Applications*, *129*(13), 28-30. https://doi.org/10.5120/ijca2015907079

Gao, Y., Gu, L., Zhou, B., Sarikaya, R., Afify, M., Kuo, H. K., ... & Besacier, L. (2006, June). IBM MASTOR SYSTEM: Multilingual automatic speech-to-speech translator. In *Proceedings of the Workshop on Medical Speech Translation* (pp. 53-56). Association for Computational Linguistics. https://doi.org/10.3115/1706257.1706268

Gates, D., Lavie, A., Levin, L., Waibel, A., Gavaldà, M., Mayfield, L., ... & Zhan, P. (1996, August). End-to-end Evaluation in JANUS: a Speech-to-speech Translation System. In *Workshop on Dialogue Processing in Spoken Language Systems* (pp. 195-206). Springer Berlin Heidelberg.

Godden, K. (2002, July). Towards a speech-to-speech machine translation quality metric. In *Proceedings of the ACL-02 workshop on Speech-to-speech translation: algorithms and systems-Volume 7* (pp. 117-120). Association for Computational Linguistics. https://doi.org/10.3115/1118656.1118672

Hamon, O., & Mostefa, D. (2008, May). An Experimental Methodology for an End-to-End Evaluation in Speech-to-Speech Translation. In *LREC*.

He, X., Deng, L., & Acero, A. (2011, May). Why word error rate is not a good metric for speech recognizer training for the speech translation task?. In *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on* (pp. 5632-5635). IEEE. https://doi.org/10.1109/icassp.2011.5947637

Levin, L., Gates, D., Lavie, A., Pianesi, F., Wallace, D., Watanabe, T., & Woszczyna, M. (2000, April). Evaluation of a practical interlingua for task-oriented dialogue. In *Proceedings of the 2000 NAACL-ANLP Workshop on Applied interlinguas: practical applications of interlingual approaches to NLP-Volume 2* (pp. 18-23). Association for Computational Linguistics. https://doi.org/10.3115/1117554.1117557

Narayanan, S., Georgiou, P. G., Sethy, A., Wang, D., Bulut, M., Sundaram, S., ... & Vergyri, D. (2006, May). Speech recognition engineering issues in speech to speech translation system design for low resource languages and domains. In *Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on* (Vol. 5, pp. V-V). IEEE. https://doi.org/10.1109/icassp.2006.1661499

Ney, H. (1999). Speech translation: Coupling of recognition and translation. In *Acoustics, Speech, and Signal Processing, 1999. Proceedings., 1999 IEEE International Conference on* (Vol. 1, pp. 517-520). IEEE. https://doi.org/10.1109/icassp.1999.758176

Nübel, R. (1997). End-to-End evaluation in VERBMOBIL I. *Proceedings of MT Summit VI*, 232-239.

Owczarzak, K., Van Genabith, J., & Way, A. (2007, April). Dependency-based automatic evaluation for machine translation. In *Proceedings of the NAACL-HLT 2007/AMTA Workshop on Syntax and Structure in Statistical Translation* (pp. 80-87). Association for Computational Linguistics. https://doi.org/10.3115/1626281.1626292

Sanders, G. A., Weiss, B. A., Schlenoff, C., Steves, M. P., & Condon, S. (2013). Evaluation methodology and metrics employed to assess the TRANSTAC two-way, speech-to-speech translation systems. *Computer Speech & Language*, *27*(2), 528-553. https://doi.org/10.1016/j.csl.2011.05.001

Sangeetha, J., & Jothilakshmi, S. (2015). Integrating machine translation and speech synthesis component for English to Dravidian language speech to speech translation system. *Journal of Engineering Science and Technology*, *10*(2), 196-211.

Tran, D. T. (2000). *Fuzzy Approaches to Speech and Speaker Recognition*. University of Canberra. https://doi.org/10.1109/nafips.1999.781728

Treichler, J. (2009). Signal processing: A view of the future, part 2 [exploratory dsp]. *IEEE Signal Processing Magazine*, *26*(3).83-86 https://doi.org/10.1109/msp.2009.932165

Wahlster, W. (Ed.). (2013). *Verbmobil: foundations of speech-to-speech translation*. Springer Science & Business Media. https://doi.org/10.1007/978-3-662-04230-4

Wang, C., & Seneff, S. (2004). High-quality speech translation for language learning. In *InSTIL/ICALL Symposium 2004*. https://doi.org/10.1145/1149290.1149291

Weiss, B. A., & Schlenoff, C. I. (2011). Performance Assessments of Two-Way, Free-Form, Speech-to-Speech Translation Systems for Tactical Use. National Institute of Standards and Technology Gaithersburg MD.

Zhang, R., Kikui, G., Yamamoto, H., Watanabe, T., Soong, F., & Lo, W. K. (2004, August). A unified approach in speech-to-speech translation: integrating features of speech recognition and machine translation. InProceedings of the 20th international conference on Computational Linguistics (p. 1168). Association for Computational Linguistics.47 https://doi.org/10.3115/1220355.1220523