INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

Capturing Formality in Speech Across Domains and Languages

Debasmita Bhattacharya[1], *Jie Chi*[2], *Julia Hirschberg*[1], *Peter Bell*[2]

[1]Columbia University, USA
[2]University of Edinburgh, UK

debasmita.b@cs.columbia.edu, jie.chi@ed.ac.uk, julia@cs.columbia.edu, peter.bell@ed.ac.uk

Abstract

The linguistic notion of formality is one dimension of stylistic variation in human communication. A universal characteristic of language production, formality has surface-level realizations in written and spoken language. In this work, we explore ways of measuring the formality of such realizations in multilingual speech corpora across a wide range of domains. We compare measures of formality, contrasting textual and acoustic-prosodic metrics. We believe that a combination of these should correlate well with downstream applications. Our findings include: an indication that certain prosodic variables might play a stronger role than others; no correlation between prosodic and textual measures; limited evidence for anticipated inter-domain trends, but some evidence of consistency of measures between languages. We conclude that non-lexical indicators of formality in speech may be more subtle than our initial expectations, motivating further work on reliably encoding spoken formality.
Index Terms: formality, speech analysis, code-switching, prosody

1. Introduction

The linguistic notion of formality can be defined as one of the many dimensions of stylistic variation, notably introduced in [1]. This particular aspect of stylistic variation is a way of characterizing interaction with others; whenever speakers interact with one another they adopt a specific *register*, i.e. level of formality. As a universal characteristic of language production [2, 3], formality has surface-level realizations in the form of written and spoken language. We are interested in measuring the formality of such realizations in order to gain a deeper understanding of the underlying linguistic phenomenon. However, characterizing formality has proven difficult from a computational perspective, as measurements of formality need to be validated, which is difficult without human-labeled data. Additionally, most existing measurement techniques have been developed for the written rather than the spoken domain.

Being able to measure formality in speech is useful, both for understanding it more clearly and for a number of downstream applications, including in the domain of code-switching. Additional possible application areas include the improvement of virtual assistants' conversational style (as alluded to in [4]), making artificially synthesized speech sound more natural and/or context-appropriate, augmenting prior work on charisma and likeability (e.g. [5]), and improving L2 language education.

A major challenge in studying spoken language formality is that it is hard to reliably characterize at the level of an individual conversation. We avoid this problem by developing an innovative validation framework operating across multiple domains or genres of language where we have a reasonable belief that average formality follows a certain trend. These can serve as a proxy for discrete labels of formality level, because certain domains of speech can, to some extent, be treated as having an inherent and stable level of formality [3]; for example, in any language, speech from broadcast news tends to be quite formal, while telephone conversations between friends tend to be relatively informal.

This work serves as a computational exploration of the linguistic notion of formality through examination of its written and spoken realizations. In the first part of this paper, we analyze formality across diverse written and spoken corpora, investigating textual and acoustic-prosodic measures of formality across languages and domains. We verify that there are surface realizations of formality that are language-independent. In particular, we find that F-score [2] and standard deviation in jitter, shimmer, speaking rate, and mean pausal duration are generally promising measures of formality in text and speech respectively. However, we find that measures of formality derived from text-based representations are not consistent with acoustic-based measures. We then explore an application of our findings on a select number of code-switched speech corpora. We believe formality to have a particularly strong association with code-switching behavior, and therefore look specifically at code-switching contexts to ascertain whether measures of formality generalize beyond monolingual contexts. We find that these do indeed generalize to code-switched contexts.

The novelty of our work is primarily based on the fact that, although the concept of formality in written language is relatively well understood, this is not the case for spoken language. We have been ambitious in conducting a multilingual study across several corpora. Our contributions include validating an existing measure of textual formality on transcripts of spoken language in multiple languages and across a wide range of domains; and moving beyond previous research in attempting to do the same for speech. Given the fundamental linguistic notion of formality, we suggest that both lexical and acoustic aspects of spoken language should encode the same patterns of formality (see Section 2). Thus, we looked for raw characteristics of speech that encode the same linguistic signals of formality as textual features do. While our findings are not definitive, they are promising and we hope that this work will encourage further study in this area.

2. Related work

Though formality has not been well-studied computationally, there has been some prior research in the area that serves as the

Part of this work was done during JSALT 2022 at JHU, with gift-funds from Amazon, Microsoft and Google. This work was also supported in part by the UKRI (grant EP/S022481/1).

10.21437/Interspeech.2023-1852

foundation of our work. The earliest work on formality was focused on characterizing the linguistic phenomenon in writing. Notably, [6] and [7] defined a five-level scale of formality and established the notion of formality in terms of syntactic and lexical differences, respectively. Similarly, [8] characterized formality as a "cohesive device" of language that manifests itself in the form of specific linguistic constructions. [9] complemented this work by proposing that formality can be understood in terms of the clarity and effort put into language production. Although there have been many proposed definitions and frameworks of written linguistic formality, and many people can intuitively distinguish between formal and informal language [10, 11, 12], "it is an ongoing challenge to grasp the exact relation between particular speech situations and their corresponding linguistic characteristics" [13].

Much previous work to address this difficulty in extracting non-lexical aspects of spoken language meaning has focused on determining the characteristics of formal English, and has shifted towards *measuring*, rather than characterizing, formality in written language. In particular, [2] proposed one of the first metrics of formality in written language, the *F-score*, based on the distribution of various parts of speech, motivated by the findings of [7]. [14] and [15] studied the relationship between formality and the use of contractions in emails and website text respectively, measuring its presence in informal communication. More recently, [16] developed models of formality perception at both the multi-textual- and word-level and compared these to existing metrics including F-score, finding that certain language, discourse, and psychological features are better than others at capturing formality as humans perceive it.

In addition to studies of written language, there has been some work examining formality in spoken language transcripts, e.g. [17, 18, 19]. Notably, [20] showed that adjective density is an appropriate indicator of formality in both transcribed speech and written language.

In contrast to studies of speech transcripts, some work has measured acoustic-prosodic and linguistic features of formal and informal speech directly. Numerous authors, including [21, 22, 23], have found that speaking rate is an important indicator of formality across Korean, Elche Spanish, and Japanese respectively. [13]'s investigation of Dutch uncovered relationships between lower articulation rates and formal interactions, as well as the greater presence of interjections, filled gaps, laughter, and disfluencies in informal interactions. In a follow-up study, [24] reinforced the idea that there are possible cross-linguistic patterns in the relationship between prosody and formality, referring to work done in Catalan in addition to previously mentioned studies in Korean and Japanese. However, each of these studies focused on a single language; in this paper we seek to establish whether we can characterize and measure formality in a way that generalizes across languages and domains.

3. Research questions

The research questions we address in this work are:

1. Do metrics that have been developed for textual realizations of formality generalize to lexical representations of speech across languages?
2. Can traditional acoustic-prosodic features of speech reliably measure formality in its spoken realization across languages?
3. Are metrics developed for text consistent with metrics developed for speech?
4. Do the answers to all of the above questions generalize beyond monolingual contexts?

It is hard to predict whether metrics developed for textual formality will hold on speech transcripts, and harder still to predict the reliability of speech features in encoding the linguistic notion of formality. But, if we uncover affirmative answers to questions 1 and 2, we expect an affirmative answer to question 3, as written and spoken language ultimately represent the same underlying linguistic notion of formality. And, if our findings generalize across multiple languages in monolingual settings, we might expect them to also generalize *between* languages within code-switched contexts.

4. Corpora

We examine a number of monolingual English, Spanish, Mandarin, and Hindi spoken corpora across a predefined set of domains, listed in Table 1. We designate written language data sets as those that contain transcripts of speech and spoken language data sets as those that contain speech audio. We choose a number of domains ranging from relatively informal telephone conversations and instructional fitness YouTube videos, to relatively formal TED Talks, broadcast news, and legal proceedings.[1]

We also consider three multilingual code-switched corpora: the Bangor Miami Spanish-English corpus [25], the SEAME Mandarin-English corpus [26], and the All-CS Hindi-English corpus.[2] The former two consist of interview-based and conversational data where speakers code-switch on occasion, while the latter consists of movie scripts.

5. Method

5.1. Written language data sets

We assume throughout this work that a corpus has an established level of formality. For the monolingual data sets, we perform part-of-speech (POS) tagging[3] using the spacy and stanza Python libraries, and use the resulting tags to compute F-scores. The formula for F-score comes from [2], whose authors proposed that *formal* language is more precise, structurally complex, and context-independent than *informal* language. This distinction is reflected in certain categories of parts of speech lending themselves to more precise and unambiguous language, and other categories producing more implicit language:

$$\begin{aligned} F = (&\text{noun frequency} + \text{adjective freq.} + \text{preposition freq.} \\ &+\text{article freq.}–\text{pronoun freq.}–\text{verb freq.} \\ &-\text{adverb freq.}–\text{interjection freq.} + 100)/2 \end{aligned} \quad (1)$$

For the code-switched data sets, we first use taggers of both relevant languages to obtain two sets of POS tags for each sentence. We then perform token-level language identification for each sentence and create a final sequence of POS tags by choosing each token's final tag based on its corresponding identified language.

[1] Self-collected YouTube and TED Talks Hindi data can be found at: `https://tinyurl.com/3vjtberc`. NOW - News on the Web data can be found at: `https://www.corpusdata.org/formats.asp`.

[2] The All-CS and All India Radio data sets were kindly made available to us by Preethi Jyothi at The Indian Institute of Technology Bombay.

[3] For Mandarin data, we additionally use the Jieba tokenizer.

Table 1: *Summary of written and spoken language data sets, in order of increasing formality.*

Speech settings	Text	Speech
Telephone conversations	CallHome (en, es) [27, 28], HUB5(zh) [29]	CallHome (en, es) [27, 28], HUB5(zh) [29], CALLFriend(hi)[30]
YouTube	YouTube(en,es,zh)	YouTube(en,es,zh,hi)
TED Talks	Multilingual TEDx (en, es) [31], Multitarget TED(en,zh) [32], TED Talks India(hi)	Multilingual TEDx (en, es) [31] TED Talks India(hi)
Broadcast news	NOW(en), HUB4-NE(es) [33], TDT4(en,zh) [34], All-India Radio(hi)	HUB4-NE(es) [33], TDT4(en,zh) [34], All-India Radio(hi)
Legal proceedings	SigmaLaw(en) [35], Europarl v7(es) [36], UN(zh) [37], IITB(hi)[38]	/
Code-switched	Bangor Miami(en-es)[25], SEAME(en-zh)[26], ALL-CS(en-hi)	Bangor Miami(en-es)[25], SEAME(en-zh)[26]

5.2. Spoken language data sets

We compute the following conversation-level acoustic-prosodic features for each corpus, separated by speaker. We examine only the spoken segments of each audio, using a silence threshold to separate speech from silence at the top-level.

5.2.1. Voice quality measures

Pitch, intensity, jitter, and shimmer are selected to represent voice quality and prosodic information in our experiments. We use Praat to extract the mean and standard deviation of pitch and intensity within each audio automatically, as well as the mean of jitter and shimmer. We set the pitch floor to 75Hz and pitch ceiling to 600Hz. For intensity, we use the 'energy' averaging method. We extract local jitter only, setting period floor to 0.0001s, period ceiling to 0.02s, and maximum period factor to 1.3. We use the same settings for local shimmer, and set maximum amplitude factor to 1.6.

5.2.2. Speaking rate

For corpora with timestamps indicating boundaries between utterances, we use these to calculate the entire duration of voiced activity. Otherwise, we use Praat to extract duration with a silence threshold of -25dB and a minimal sounding interval of 0.1s. We then calculate the average number of words per second, as well as the number of words per utterance.

5.2.3. Pauses

Similar to speaking rate, we first obtain the total duration of the utterances and use Praat to extract pauses, where a pause is defined as a silent interval of less than 2s. We then calculate both the average duration per pause and the number of pauses per utterance.

6. Results

6.1. F-score generalizes to lexical representations of speech across languages

We begin by validating F-score as a measure of written formality. We find that the metric seems to generalize across languages and domains as its value generally increases in more formal domains, as expected. Figure 1 additionally reveals that within a given genre, Hindi generally scores the highest, followed by Spanish, then English, and finally Mandarin. We expect this trend is a reflection of the syntactic complexity of each of the languages under consideration, as the F-score draws directly on part-of-speech information.

6.2. Variation in speaking rate, jitter, shimmer, and mean pausal duration may encode formality across languages

We next turn to possible measures of spoken formality. We focus on the most promising corpus-level speech features of those that we examined — i.e. the ones that showed a generally consistent effect: standard deviation in speaking rate, jitter, and shimmer, and mean pausal duration (Figure 2).

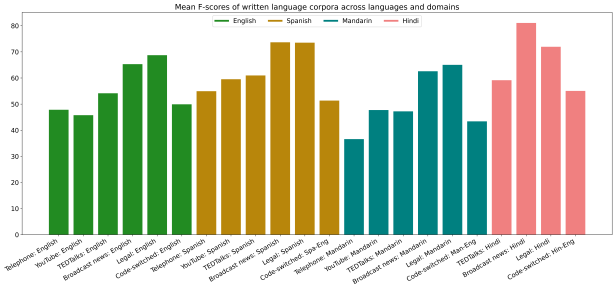

Figure 1: *Comparing F-scores of written language corpora across domains and languages. Mean values within domains across all four languages exhibit the same patterns as above.*

First, we find some general patterns across domains and languages in standard deviation in speaking rate, measured as words per second. In English, we see some indication of greater informality being encoded by greater variation in speaking rate, as suggested by telephone conversations having a higher standard deviation value than TED Talks. Similarly, in Spanish, telephone conversations have the greatest standard deviation, followed by TED Talks, and finally broadcast news. But, we do not find any such coherent patterns in the Mandarin or Hindi data. We also note that the instructional fitness YouTube corpora do not behave as expected, leading us to conclude that we were not entirely successful in controlling for inherent variation in the domain. We ignore the instructional fitness YouTube domain in the remainder of our discussion.

The next two promising features that we uncover in our exploration are standard deviation in jitter and shimmer. In English, as above, we find a similar pattern in standard deviation in jitter and shimmer, where telephone conversations have greater values than TED Talks. We find comparable patterns in Spanish, with broadcast news having the lowest standard deviation across domains. In Hindi, too, telephone conversations have greater standard deviation in jitter and shimmer than TED Talks. However, while broadcast news has the lowest standard deviation in shimmer across genres, it is odd that the domain has higher standard deviation in jitter than TED Talks.

The final promising feature we find is the mean duration per pause. In English, the ordering of (informal) telephone conversations above more prepared TED Talks aligns with our previously found patterns. In Spanish, broadcast news having the lowest mean pausal duration also makes sense, as above; however the remaining domains are similar to one another, making any distinction between them difficult to find. In Hindi, telephone conversations have the highest mean pausal duration and broadcast news has the lowest value, with TED Talks in between; these results align with our expectations.

Overall, the patterns we find in this section are similar to, but less consistent than, those we uncovered in the written do-

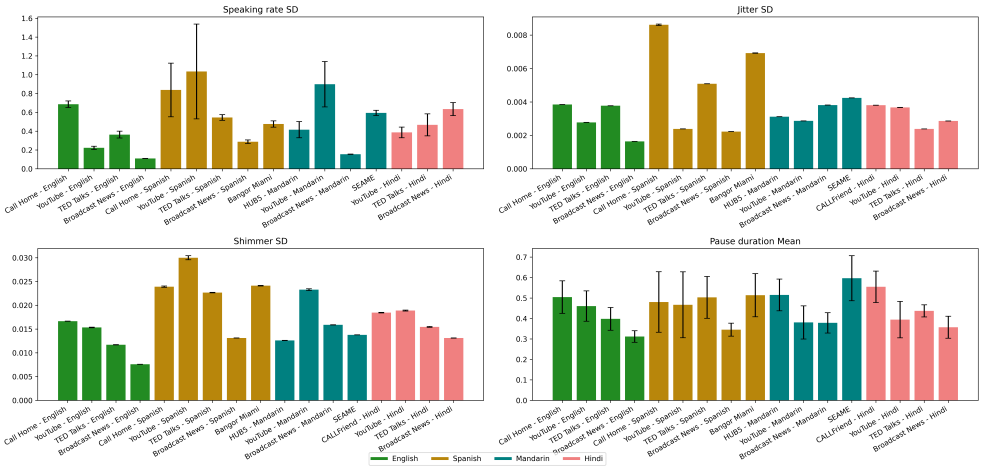

Figure 2: *Comparing speaking rate, jitter, shimmer and pause duration across spoken language corpora.*

mains. However, our findings largely support previous work (e.g. [23, 24]) that has pointed to the importance of *variation* in speech variables in encoding formality of spoken language.

6.3. Textual measures of formality do not seem consistent with spoken measures

We perform correlation analysis at the domain- and language-level, treating each conversation as a data point. We generally do not find any correlation between F-score and any of the speech variables mentioned in the previous section. R-squared values from each of these attempted correlations have a minimum of 0.000, maximum of 0.644, mean of 0.034, median of 0.015, and mode of 0.000. We had expected to find some correlation between text-based measures of formality and speech-based measures as we believe that both written and spoken surface realizations of formality ultimately reflect the same underlying phenomenon. We know that humans are good at judging formality [10, 11], so the lack of correlation tells us that the way in which the linguistic notion of formality manifests itself in the prosodic features of speech is more subtle than expected.

6.4. Patterns in monolingual data generalize to code-switched contexts

Finally, we examine our code-switched corpora and check whether the patterns that appear on monolingual data generalize to code-switched data. Note that we continue to ignore the instructional fitness YouTube domain in the following discussion.

We largely find that the patterns do generalize. First, the mean F-score for code-switched data is similar to that for telephone conversations, the lowest (most *informal*) among the domains under consideration. Further, across standard deviation in speaking rate, jitter, and shimmer, and mean pausal duration, values for the code-switched Spanish-English and Mandarin-English corpora are either greater than or very similar to those for telephone conversations, the most informal of the monolingual domains. These patterns are what we would expect since the code-switched data sets are also conversational and are probably about the same level of formality as the telephone conversation data. One exception we find is in standard deviation in speaking rate for the Spanish-English data, where the code-switched domain has a lower value than that of Spanish TED Talks, indicating a higher level of formality than in TED Talks. We think that the Spanish TED Talks data may not have been representative of the domain due to large variations in duration (as short as 2 minutes) and nature (including singing and other musical performance) of talks. Overall, we find indications that textual and spoken measures of formality generalize beyond monolingual contexts to code-switched ones.

We also look at the correlation between text and speech measures for the code-switched corpora. As with the monolingual corpora, we find no correlation between conversation-level F-score and any of the relevant speech variables, further supporting our conclusion that patterns of formality in monolingual data generalize to code-switched contexts.

7. Conclusion

In this work, we measure the formality of surface-level realizations of speech. In response to the four research questions we posed at the outset, we find: (1) metrics that have been developed for textual realizations of formality generalize to lexical representations of speech, with this result being consistent across the languages we investigated; (2)(a) an initial indication that certain traditional acoustic-prosodic variables might play a stronger role than others in reliably encoding formality in speech, though we find limited evidence for our anticipated inter-domain trends; (2)(b) some evidence of consistency of speech-based measures between the languages investigated; (3) no correlation between text-based measures and speech-based measures; and (4) patterns in monolingual domains are also reflected in code-switched domains.

We conclude that non-lexical indicators of formality may be more subtle than our initial expectations. While in written domains the measurement of formality is quite successful, on the spoken side this remains a challenging task. Further work is required to determine how strong the effect of our promising speech variables is and to definitively produce a reliable formality score from audio features of spoken language. It may be fruitful to consider additional speech variables, e.g. speaking rate measured in terms of syllables and the duration of vowels in particular, or to consider interaction effects of current features on formality. Another inspiring direction to pursue might involve using off-the-shelf acoustic representations that have conventionally been used in automatic speech recognition and are starting to be used in prosodic analysis [39], but it is an open question as to how we would use such representations in the absence of gold standard labels for formality, an inherent limitation of study in this area.

8. References

[1] W. Labov, *Sociolinguistic patterns*. University of Pennsylvania Press, 1972.

[2] F. Heylighen and J. Dewaele, "Formality of Language: definition, measurement and behavioral determinants," 1999.

[3] L. M. Tanaka, "The strategic use of speech style shifts in Japanese radio." 2008.

[4] M. de Jong, M. Theune, and D. Hofs, "Politeness and alignment in dialogues with a virtual guide," vol. 1, 01 2008, pp. 207–214.

[5] A. Rosenberg and J. Hirschberg, "Charisma perception from text and speech," *Speech Communication*, vol. 51, no. 7, pp. 640–655, 2009.

[6] M. Joos, *The Five Clocks*. Harcourt Brace, 1962.

[7] P. E. Brown and C. Fraser, "Speech as a marker of situation," 1979.

[8] G. Cook, *Discourse in language teaching: A scheme for teacher education.* Oxford University Press, 1989.

[9] D. Biber and E. Finegan, *Sociolinguistic Perspectives on Register*. Oxford University Press, 1994.

[10] C. Creber and H. Giles, "Social context and language attitudes: The role of formality-informality of the setting," *Language Sciences*, vol. 5, pp. 155–161, 1983.

[11] S. Lahiri, P. Mitra, and X. Lu, "Informality Judgment at Sentence Level and Experiments with Formality Score," in *Computational Linguistics and Intelligent Text Processing*. Berlin, Heidelberg: Springer, 2011, pp. 446–457.

[12] S. Wallbridge, P. Bell, and C. Lai, "It's Not What You Said, it's How You Said it: Discriminative Perception of Speech as a Multichannel Communication System," in *Proc. Interspeech 2021*, 2021, pp. 2386–2390.

[13] K. Koppen, M. Ernestus, and M. van Mulken, "The influence of social distance on speech behavior: Formality variation in casual speech," *Corpus Linguistics and Linguistic Theory*, vol. 15, 01 2017.

[14] E. Turney, C. Sabater, and B. Montero-Fleta, "Formality and informality in electronic communication," 08 2006, pp. 241–244.

[15] S. Karlsson, "Formality in Websites: differences regarding country of origin and market sector," 2008.

[16] H. Li, A. Graesser, M. Conley, Z. Cai, P. Pavlik Jr, and J. Pennebaker, "A New Measure of Text Formality: An Analysis of Discourse of Mao Zedong," *Discourse Processes*, vol. 53, pp. 0–0, 02 2015.

[17] H. Levin and P. Garrett, "Sentence structure and formality," *Language in Society*, vol. 19, no. 4, pp. 511–520, 1990.

[18] K. B. Dempsey, P. M. McCarthy, and D. S. McNamara, "Using Phrasal Verbs as an Index to Distinguish Text Genres," in *The Florida AI Research Society*, 2007.

[19] R. Rittman, "Automatic discrimination of genres: The role of adjectives and adverbs as suggested by linguistics and psychology," 2007.

[20] A. C. Fang and J. Cao, "Adjective Density as a Text Formality Characteristic for Automatic Text Classification: A Study Based on the British National Corpus," in *Proceedings of the 23rd Pacific Asia Conference on Language, Information and Computation, Volume 1*. Hong Kong: City University of Hong Kong, Dec. 2009, pp. 130–139. [Online]. Available: https://aclanthology.org/Y09-1015

[21] B. Winter and S. Grawunder, "The phonetic profile of Korean formal and informal speech registers," *Journal of Phonetics*, vol. 40, p. 808–815, 11 2012.

[22] W. Chappell, "Casual speech or fast speech?: A qualification about the effect of formality and speech rate on Spanish /s/ reduction," *International Journal of the Linguistic Association of the Southwest*, vol. 33, 12 2014.

[23] E. Sherr-Ziarko, "Acoustic Properties of Formality in Conversational Japanese," in *Interspeech*, 2016.

[24] ——, "Prosodic properties of formality in conversational Japanese," *Journal of the International Phonetic Association*, vol. 49, no. 3, p. 331–352, 2019.

[25] M. Deuchar, "Miami corpus: Preliminary documentation - bangortalk." [Online]. Available: http://bangortalk.org.uk/docs/Miami_doc.pdf

[26] D.-C. Lyu, T.-P. Tan, E. Chng, and H. Li, "Mandarin–English code-switching speech corpus in South-East Asia: SEAME," vol. 49, 01 2010, pp. 1986–1989.

[27] A. Canavan, D. Graff, and G. Zipperlen, "CALLHOME American English Speech," 1997.

[28] B. Wheatley, "CALLHOME Spanish Transcripts LDC96T17," 1996.

[29] L. D. Consortium, "HUB5 Mandarin Telephone Speech and Transcripts Second Edition LDC2018S18," 2018.

[30] A. Canavan and G. Zipperlen, "CALLFRIEND Hindi LDC96S52," 1996.

[31] E. Salesky, M. Wiesner, J. Bremerman, R. Cattoni, M. Negri, M. Turchi, D. W. Oard, and M. Post, "The Multilingual TEDx Corpus for Speech Recognition and Translation," *CoRR*, vol. abs/2102.01757, 2021. [Online]. Available: https://arxiv.org/abs/2102.01757

[32] K. Duh, "The Multitarget TED Talks Task," http://www.cs.jhu.edu/~kevinduh/a/multitarget-tedtalks/, 2018.

[33] S. Huang, J. Liu, X. Wu, L. Wu, Y. Yan, and Z. Qin, "1997 Mandarin Broadcast News Transcripts (HUB4-NE) LDC98T24," 1998.

[34] J. Kong and D. Graff, "TDT4 Multilingual Broadcast News Speech Corpus LDC2005S11."

[35] C. R. Mudalige, D. Karunarathna, I. Rajapaksha, N. de Silva, G. Ratnayaka, A. S. Perera, and R. Pathirana, "SigmaLaw-ABSA: Dataset for Aspect-Based Sentiment Analysis in Legal Opinion Texts," *CoRR*, vol. abs/2011.06326, 2020. [Online]. Available: https://arxiv.org/abs/2011.06326

[36] P. Koehn, "Europarl: A Parallel Corpus for Statistical Machine Translation," in *Proceedings of Machine Translation Summit X: Papers*, Phuket, Thailand, Sep. 13-15 2005, pp. 79–86. [Online]. Available: https://aclanthology.org/2005.mtsummit-papers.11

[37] M. Ziemski, M. Junczys-Dowmunt, and B. Pouliquen, "The United Nations Parallel Corpus v1.0," in *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*. Portorož, Slovenia: European Language Resources Association (ELRA), May 2016, pp. 3530–3534. [Online]. Available: https://aclanthology.org/L16-1561

[38] A. Kunchukuttan, P. Mehta, and P. Bhattacharyya, "The IIT Bombay English-Hindi parallel corpus," in *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. Miyazaki, Japan: European Language Resources Association (ELRA), May 2018. [Online]. Available: https://aclanthology.org/L18-1548

[39] G.-T. Lin, C.-L. Feng, W.-P. Huang, Y. Tseng, T.-H. Lin, C.-A. Li, H.-y. Lee, and N. G. Ward, "On the utility of self-supervised models for prosody-related tasks," in *2022 IEEE Spoken Language Technology Workshop (SLT)*, 2023, pp. 1104–1111.