Interspeech 2018
2-6 September 2018, Hyderabad, India

Data Augmentation Improves Recognition of Foreign Accented Speech

Takashi Fukuda[1], *Raul Fernandez*[2], *Andrew Rosenberg*[2], *Samuel Thomas*[2],
Bhuvana Ramabhadran[†], *Alexander Sorin*[3], *Gakuto Kurata*[1]

[1,2,3]IBM Research AI
[†]Google

fukuda1@jp.ibm.com, {fernanra, amrosenb, sthomas}@us.ibm.com,
bhuv@google.com, sorin@il.ibm.com, gakuto@jp.ibm.com

Abstract

Speech recognition of foreign accented (non-native or L2) speech remains a challenge to the state-of-the-art. The most common approach to address this scenario involves the collection and transcription of accented speech, and incorporating this into the training data. However, the amount of accented data is dwarfed by the amount of material from native (L1) speakers, limiting the impact of the additional material. In this work, we address this problem via data augmentation. We create modified copies of two accents, Latin American and Asian accented English speech with voice transformation (modifying glottal source and vocal tract parameters), noise addition, and speed modification. We investigate both supervised (where transcription of the accented data is available) and unsupervised approaches to using the accented data and associated augmentations. We find that all augmentations provide improvements, with the largest gains coming from speed modification, then voice transformation and noise addition providing the least improvement. The improvements from training accent specific models with the augmented data are substantial. Improvements from supervised and unsupervised adaptation (or training with soft labels) with the augmented data are relatively minor. Overall, we find speed modification to be a remarkably reliable data augmentation technique for improving recognition of foreign accented speech. Our strategies with associated augmentations provide Word Error Rate (WER) reductions of up to 30% relative over a baseline trained with only the accented data.

Index Terms: speech recognition, data augmentation, voice transformation, foreign accented speech

1. Introduction

The state of the art in automatic speech recognition (ASR) performance continues to improve, in some cases approaching human levels of performance. However, the recognition of foreign accented speech is still a challenge, with performance significantly lagging. For a global language like English, estimates suggest that non-native speakers may outnumber native speakers by as much as 3 to 1 [1]. This is a well-known problem, and there are a variety of ways that have been explored to address it (cf. Section 2). Most of these approaches involve collecting some amount of accented speech and using it for training either instead of or alongside the available native speaker data.

Despite the large population of non-native speakers of English, there is much more high quality speech data from native speakers than non-native speakers available in the form of publicly available and even most privately held data sets. Moreover, non-native speech is a brief descriptor of an incredibly heterogeneous set. A particular speaker's specific native language, and their experience and proficiency in speaking English all have significant impacts on the realization of their speech. Novice learners of English speak differently than fluent speakers; Native Japanese speakers speak English differently than Native Spanish speakers do. These are dimensions of variation that are unique to foreign accented data, and operate in addition to the variation that impacts recognition of native English speech, like speaker differences and recording conditions. This variation compounds the problem of available data – not only is there less foreign accented data available, but it is more varied than the corresponding native data.

A common approach to address this limitation is data augmentation (Section 3), wherein artificial copies of the available audio data are generated using a label-preserving transformation. The model topology, training and adaptation methods used for this work are presented in Section 4. The main contributions of this work are:

- A novel, voice transformation technique based on acoustic source- and vocal-tract transformations as a data augmentation scheme.
- Impact of three approaches to data augmentation: voice transformation, noise addition and speed modification in supervised settings when training models anew and via adaptation (Section 5.2).
- Impact of using foreign accented data for unsupervised adaptation and the value of the aforesaid data augmentation methods in this context (Section 5.3).

Unsurprisingly, we find that in both the supervised and unsupervised contexts, voice transformation and speed modification based data augmentation helps with the recognition of foreign accented speech. More surprisingly, despite being the simplest transformation approach, we find that speed modification is a remarkably robust data augmentation approach, yielding larger gains than voice transformation or noise addition when dealing with recognition of accented speech.

2. Related Work

Data augmentation is a commonly employed technique to increase the diversity of training data by artificially creating additional training samples. Increasing the amount of training data using various signal or data processing techniques has shown to be consistently beneficial by preventing models from over

[1]IBM Tokyo Research Labs, Tokyo, Japan
[2]IBM T. J. Watson Research Center, Yorktown Heights, NY
[3]IBM Haifa Research Labs, Haifa, Israel
[†]Work performed while at IBM.

The authors wish to thank Kartik Audhkhasi and Slava Schectman for useful discussions.

10.21437/Interspeech.2018-1211

fitting and improving the overall robustness of ASR models. In [2] data augmentation by creating multiple versions of the original signal with various speed factors has been shown to improve ASR performance across various tasks. This approach is used more elaborately in [3] where in addition to data augmentation at the signal level with noise addition and speed perturbation, data is augmented in a second stage, at the feature level with a fMLLR-based technique applied to bottleneck features. Similarly in [4], data augmentation is performed at the feature level via vocal tract length perturbation (VTLP) and stochastic feature mapping (SFM), to improve ASR performance in low resource settings. Data augmentation has also been investigated via semi-supervised training, speech synthesis and multi-lingual processing [5]. In most ASR settings since noise robustness is desired, data augmentation by corrupting clean train data with various additive and convolutive noises has also been extensively studied [6].

In this work, we investigate the usefulness of several of these data augmentation techniques at the *audio signal level* for recognition of accented speech. We introduce a novel, acoustic source- and vocal-tract transformations as the basis of a data-augmentation technique. This technique was originally proposed for a prosody-labeling task [7]. Its use to enhance acoustic modeling for speech recognition and speed modification focusing on accented speech is, to the best of our knowledge, a novel contribution of this paper.

3. Data Augmentation Strategies

In this section we describe three approaches to data augmentation: Voice Transformation (Section 3.1), Noise Addition (Section 3.2) and Speed Modification (Section 3.3).

3.1. Voice Transformation

In this work we explore the application of a voice-transformation technique capable of manipulating the vocal-source and vocal-tract characteristics to alter the speaker's voice quality and/or impart novel speaker identities. Since the modifications preserve the phonetic content of utterances, they may be paired with the original transcripts to create additional training data in a supervised-learning approach (Section 5.2). In this section we offer a brief overview of the analysis and resynthesis technique; a more detailed exposition can be found in [8].

Analysis: A pitch contour is first extracted from the audio signal at a 5ms-update rate, and the signal analyzed at the same frame rate. The analysis is limited to voiced regions (all unvoiced frames are skipped), with a short-time window of 3.5 pitch cycles. The Iterative Adaptive Inverse Filtering algorithm [9] is employed to obtain a raw glottal source derivative waveform, and this one fitted with the Liljencrants-Fant (LF) parametrization [10] represented by a 3-parameter vector $\theta = [T_p, T_e, T_a]^T$ (normalized by the pitch period), the aspiration noise level, and the gain factor. The vocal tract is represented by 40 Line Spectral Frequencies (LSF). The temporal trajectories of all parameters are smoothed with a 7-frames long moving averaging window to reduce estimation artifacts.

Resynthesis: To reconstruct the audio signal from its parameters, consecutive voiced frames are stacked together to form contiguous voiced regions, and then interleaved with unvoiced regions, which are kept in the raw form and never modified. To synthesize a consecutive voiced region, a sequence of pitch-cycle onsets is generated according to a desired synthesis F0 contour (original or transformed).

The glottal-source and vocal-tract parameters associated with each pitch cycle are generated by interpolating between the corresponding parameters associated with the cycle's surrounding (edge) frames. The sequence of glottal pulses thus generated is then multiplied by the corresponding gain factors. Additive aspiration noise is constructed for the entire voiced region by amplitude modulation of a 500-Hz high-passed Gaussian noise signal. The amplitude modulation forms the noise time-envelope shape, so that it is aligned with the glottal-pulse energy envelope, respects the noise level and gain values within each cycle, and evolves smoothly at the transitions between the consecutive cycles. The LSF parameters associated with each pitch cycle are converted into auto-regression coefficients and used to filter the pulse sequence: the filter coefficients are updated at the beginning of each pitch cycle, and each voiced region is then combined with its neighboring unvoiced regions using an overlap-add process.

Transformation: The previous algorithm provides the basis for introducing global (time-invariant) voice modifications that alter pitch, vocal tract and the glottal pulse. Global *pitch modifications* transpose and stretch the original pitch contour by factors f_0^{shift} and f_0^{range} respectively. The *vocal tract transformation* takes the form of an interpolating spline function, with user-specified inflection points, that is used to map each cycle's LSFs prior to reconstruction. For the *glottal pulse transformations* two independent types of control are supported: (i) interpolation between the actual and user-provided reference glottal-pulse vector θ_{ref} and mixing weight $0 \leq \alpha \leq 1$:

$$\hat{\theta} = (1 - \alpha)\theta + \alpha\theta_{ref} \tag{1}$$

and (ii) interpolation between two stylized anchor pulses corresponding to voice qualities that can be described as lax and tense:

$$\hat{\theta} = \begin{cases} (1 - \beta_{lt})\theta + \beta_{lt}\theta_l & \text{if } \beta_{lt} > 0 \\ (1 - |\beta_{lt}|)\theta + |\beta_{lt}|\theta_t & \text{otherwise} \end{cases} \tag{2}$$

where $-1 \leq \beta_{lt} \leq 1$ is a user-specified parameter that trades between lax and tense qualities (and recovers the original pulse when $\beta_{lt} = 0$), and $\theta_l = [.5, .9, .099]^T$ and $\theta_t = [.1, .15, .00001]^T$ are the fixed lax and tense glottal parameters respectively.

3.2. Noise Addition

As described earlier, to increase the noise robustness of ASR models, the original clean acoustic model training data is often corrupted with various additive and convolutive noises to create a multi-condition training set. A popular data set that demonstrates the usefulness of such multi-condition training is the Aurora 4 corpus [11]. In this data set, six additive noise conditions collected from street traffic, train stations, cars, babble, restaurants and airports are added to the speech data in addition to capturing convolutive noise effects via various microphone distortions. In a similar spirit, to understand the effect of noise addition as a strategy for data augmentation for accented speech, we use the FaNT (Filtering and Noise-adding Tool) tool, to create augmented data by adding noise to speech recordings at a desired SNR (signal-to-noise ratio) along with other desired frequency characteristics.

3.3. Speed Modification

Following an approach introduced in Ko et al. [2] and used by Hartmann et al. [3], we perform speed perturbation to generate modified copies of the source audio. This approach modifies the speed of each file by a multiplicative factor drawn at random

uniformly between 0.9 and 1.1. This resampling, performed using the SOX utility, impacts the duration of the file, as well as the pitch and spectral frequencies of the contained audio.

4. Models and Training

4.1. Model Topology

All acoustic models used throughout this paper are Convolutional Neural Networks (CNNs) of the same size. The CNN is trained with 40 dimensional log Mel-frequency spectra augmented with Δ and $\Delta\Delta$ as inputs. Each frame of speech is also appended with a context of 11 frames after applying a speaker independent global mean and variance normalization. The CNN system uses two convolutional layers with 128 and 256 hidden nodes each in addition to four fully connected layers with 2048 nodes per layer to estimate posterior probabilities of 9300 context-dependent states as output targets. All of the 128 nodes in the first feature extracting layer are attached with 9×9 filters that are two dimensionally convolved with the input log Mel-filterbank representations. The second feature extracting layer with 256 nodes has a similar set of 3×4 filters that processes the non-linear activations after max pooling from the preceding layer. The non-linear outputs from the second feature extracting layer are then passed onto the subsequent fully connected layers. All the layers use the ReLU non-linearity.

4.2. Weight Decay Based Supervised Adaptation

The supervised adaptation algorithm used in this paper is similar to the one proposed in [12]. This scheme resembles MAP adaptation, with the adapted weight updates arrived at from using a weighted combination of the updates from adaptation data and the baseline model. Unlike the work in [12] where adaptation was performed at a speaker level, in this paper, the entire adaptation data is pooled and the algorithm is used as an overall domain adaptation scheme [13], as given by Equation (3)

$$\Delta \mathbf{w}_t = -\alpha \nabla_{\mathbf{w}} E(\mathbf{w}_t) - \beta(\mathbf{w}_{t-1} - \mathbf{w}_0), \qquad (3)$$

where α is a learning rate, β is a regularization parameter, $E(\mathbf{w})$ is an error function, and $\mathbf{w}_0$ is model parameters of the initial model. The network is adapted using the cross entropy training criterion.

4.3. Unsupervised Adaptation Based on Teacher Student Training

A teacher student training is a framework to mimic powerful and complicated teacher networks with compact and simple student network (e.g. [14, 15]). Instead of using the ground truth labels, the teacher-student training approach defines the loss function as

$$\mathcal{L}(\theta) = -\sum_i q_i \log p_i, \qquad (4)$$

where q_i is the soft label of the teacher model, which works as a pseudo label. p_i is output probability of the class of the student model. In q_i, the competing classes will have small but nonzero posterior probabilities for each training example. Once we train powerful teachers such as VGG network, we can create student networks with lower computational complexity to approximate their performance [16, 17, 15]. In the teacher student framework, since soft labels generated from teacher networks are used as targets to train student networks, corresponding transcriptions are not always necessary. This means that teacher-student training can be leveraged as an unsupervised acoustic model training/adaptation [15] method. In this paper, we use this scheme to train as well as adapt CNN student networks in an unsupervised fashion.

5. Experiments and Results

5.1. Data

The accented data (AD) used for data augmentation is an English corpus comprising of 42.8 hours recorded from speakers with Latin American (20.7 hours from 94 speakers) and Asian (22.1 hours from 96 speakers) accents under clean and noisy conditions. A total of 38 hours from this set was used for training and the remaining (5 hours) was set aside as held-out. The test data contained 5 hours, consisting of 2.1 hours of Latin American (LA) accented speech and 2.4 hours of Asian accented speech. The utterances in this corpus are a mix of digits and alphabets in insolation, command phrases and short dialogs seen in spoken language systems. For voice-transformation based data augmentation, a total of 7 different transformations of the AD data, empirically chosen so as to provide a good amount of variability in terms of identity and expressiveness were implemented to obtain initial augmentations of the data. The augmented data was decoded with an ASR system that was not trained using the AD data. The three voice transformations with WERs less than 50% were identified as candidates for data augmentation and resulted in approximately 114 hours (VT). For speed-based data augmentation, we altered the speed of each utterance in the AD corpus by a multiplicative factor drawn at random uniformly between 0.9 and 1.1 and created 3 copies per utterance, totaling approximately 114 hours (Speed). For noise-based data augmentation, we use 12 noises from the DEMAND database [18] which includes noises in restaurants, home environments, open spaces, meeting rooms and transit modes likes buses, cars and trains. Each of the 12 noises are used to renoise approximately 10 hours of the clean portion of the AD corpus to create a noisy accented corpus of approximately 120 hours (Noise). The decoder uses a vocabulary comprising 250K words and the language model is a 4-gram LM with 200M n-grams.

The training data for the baseline model used in the adaptation experiments consists of 3600 hours of audio data. One-third of this training corpora is clean audio from three public corpora - 420 hours from broadcast news, 280 hours from Mixer 6 [19], and 100 hours from the AMI corpus [20] and 450 hours of private speech data. This corpora is further augmented with realistic environmental noises from the JEIDA corpus [21] and impulse responses from RWCP [22] at various SNRs between 5 to 20 dB to total 3600 hours.

5.2. Supervised Results

We first describe experiments where we train acoustic models *only* on the AD corpus. These accent-specific CNN models serve as a baseline (AD Baseline) and conform to the topology described in Section 4.1. The CNNs are trained with random initialization. While this AD data may be more consistent with the evaluation data, there is significantly less training data (approx. 20 hours) compared to modern ASR systems and even the baseline systems use for adaptation. If we assume accented speech in a minor language with transcriptions, this is a realistic data size. In Table 1, we compare the performance of these models with those trained with the three styles of data augmentation. Results are presented separately on the LA and Asian portion of the AD corpus.

Both Voice Transformation and Speed Modification augmentation schemes provide significant gains for both types of

Table 1: *Performance of models trained in an supervised fashion with random initialization (WER)*

Augmentation Scheme	LA Training and test WER	Asian Training and test WER
AD Baseline	23.20	26.18
AD+Noise	29.75	31.64
AD+VT	19.13	18.99
AD+Speed	17.57	18.00

Table 2: *Performance of supervised adaptation with different data augmentation schemes (WER)*

Augmentation scheme	LA supervised adaptation WER	Asian supervised adaptation WER
Unadapted	28.30	26.70
AD	14.25	13.69
AD+Noise	15.42	14.95
AD+VT	14.22	13.85
AD+Speed	14.06	13.29

accents. However, the gains are not solely a function of having more training data available, as the noise augmentation does not improve performance. It is worth noting that noise addition is a very effective data augmentation strategy to improve robustness to noise but it does not help with foreign accent performance.

Next we describe results from adapting the baseline models which are of the same topology defined in Section 4.1 and trained with the data described in Section 5.1. This is a strong unadapted baseline model that has not seen any of the AD data. Two sets of adapted models are created, one for each of the accents in the AD corpus using supervised adaptation. Adaptation is performed via weight decay as described in Section 4.2. Results are presented in Table 2.

Here we find an important improvement from adapting to the non-native data. The improvements from additional augmentations of the data based on the speed modification range 1 - 3% relative improvements over the AD data only. Once again, we find the noise-based addition to degrade performance, VT provides a very small improvement to Latin American accented speech, and a minor degradation to Asian accented speech.

5.3. Unsupervised Results

In this section, we describe results from adapting the baseline model (same as the one used in Section 5.2) using the scheme described in Section 4.3 using unsupervised adaptation. Specifically, two sets of models are created, one for each of the accents in the AD corpus using un-supervised adaptation. The teacher network used to generate the soft labels for adaptation is also trained on the same 3600 hours of data used to train the baseline model and has not seen any of the accented data. However, it is a more complex VGG model comprising 10 convolutional layers, with a max-pooling layer inserted after every 3 convolutional layers, followed by 4 fully connected layers. All hidden layers have ReLU non-linearity. Batch normalization is applied to the fully connected layers. Posteriors of the top 50 most likely labels for each prediction of the teacher are then used to adapt the baseline CNN networks. The KL-divergence criterion used for training the student model is equivalent to minimizing the cross entropy of the soft target labels. The baseline model is also adapted with the data from two augmentation schemes in the same unsupervised fashion. We chose to drop the noise-based augmentation scheme for these experiments as there was no gain with supervised adaptation (See Table 2). The results are tabulated in Table 3. We can see that the data from VT or Speed-based augmentation schemes consistently improves the performance over the system trained without any data augmentation (2.8 – 4.0% relative improvements).

Table 3: *Performance of unsupervised adaptation with different data augmentation schemes (WER)*

Augmentation scheme	LA unsupervised adaptation WER	Asian unsupervised adaptation WER
Unadapted	28.30	26.70
AD	24.75	22.03
AD+VT	23.74	21.68
AD+Speed	23.89	21.37

5.4. Leveraging Multiple Adaptation Styles

Based on the results in Section 5, we have identified that Voice Transformation and Speed Modification based data augmentation strategies are powerful techniques to improve recognition of foreign accented speech, with Speed Modification being more effective. In this section, we explore ways to determine if the improvements are complementary, and if so, how to leverage them to achieve better performance. We focus our attention on the models trained from scratch (See Table 1) and try to close the performance gap to the supervised adaptation approach (See Table 2)

Merged Training: We evaluated a system trained with half of the available VT data (57 hours) and half of the Speed Modification data (57 hours). We use only half of each set so that each system is trained on the same amount of data. Here we find the performance to be worse than augmenting using Speed Modification with a WER on LA accented speech of 18.24% and Asian accented speech, 19.69%.

Posterior Combination: Next, we combine posteriors from both systems with a 75/25 weighting favoring the AD+Speed system. Here we find that we can reduce the WER on LA accented speech further to 17.34% and that of Asian accented speech to 17.57%.

While we were able to identify an approach by which multiple data augmentation strategies can improve performance, Speed Modification remains a remarkably effective data augmentation approach on these experiments.

6. Conclusion

Data augmentation improves recognition of Latin American and Asian accented speech significantly. We have evaluated this in three contexts, unsupervised adaptation in a teacher/student framework, supervised adaptation and training acoustic models from scratch. In each of these, we found that 1) Speed Modification is a surprisingly effective data augmentation strategy, 2) Voice Transformation also helps, but is less effective, and 3) Noise Addition is the least effective and can even degrade performance. These experiments were all performed on accent-specific models, where we assume that the accent of a speaker is known ahead of time. The improvements from training accent specific models with the augmented data are substantial. Our strategies with associated augmentations provide Word Error Rate (WER) reductions of up to 30% relative over a baseline trained with only the accented data. We can see improvements from supervised and unsupervised adaptation on the augmented data as well. However the improvements are relatively minor compared to the case where we train from scratch. Overall, we find speed modification to be a remarkably reliable data augmentation technique for improving recognition of foreign accented speech.

7. References

[1] D. Crystal, *English as a Global Language*. Cambridge University Press, July 2003.

[2] T. Ko, V. Peddinti, D. Povey, and S. Khudanpur, "Audio augmentation for speech recognition." in *INTERSPEECH*, 2015, pp. 3586–3589.

[3] W. Hartmann, T. Ng, R. Hsiao, S. Tsakalidis, and R. M. Schwartz, "Two-stage data augmentation for low-resourced speech recognition," in *INTERSPEECH*, 2016, pp. 2378–2382.

[4] X. Cui, V. Goel, and B. Kingsbury, "Data augmentation for deep neural network acoustic modeling," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 23, no. 9, pp. 1469–1477, 2015.

[5] A. Ragni, K. M. Knill, S. P. Rath, and M. J. Gales, "Data augmentation for low resource languages," in *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.

[6] L. Deng, A. Acero, M. Plumpe, and X. Huang, "Large-vocabulary speech recognition under adverse acoustic environments." in *INTERSPEECH*, 2000, pp. 806–809.

[7] R. Fernandez, A. Rosenberg, A. Sorin, B. Ramabhadran, and R. Hoory, "Voice-transformation-based data augmentation for prosodic classification," in *Proc. ICASSP*, New Orleans, Louisiana, USA, March 2017, pp. 5530–5534.

[8] A. Sorin, S. Shechtman, and A. Rendel, "Semi parametric concatenative TTS with instant voice modification capabilities," in *Interspeech*, Stockholm, Sweden, August 2017, pp. 1373–1377.

[9] P. Alku, "Glottal wave analysis with pitch synchronous iterative adaptive inverse filtering," *Speech Communication*, vol. 11, pp. 109–118, 1992.

[10] G. Fant, J. Liljencrants, and Q. Lin, "A four-parameter model of the glottal flow," *STL-QPSR*, vol. 26, no. 4, pp. 1–13, 1985.

[11] D. Pearce and J. Picone, "Aurora working group: DSR front end LVCSR evaluation au/384/02," *Inst. for Signal & Inform. Process., Mississippi State Univ., Tech. Rep*, 2002.

[12] H. Liao, "Speaker adaptation of context dependent deep neural networks," *Proc. IEEE ICASSP*, pp. 7947–7951, 2013.

[13] M. Suzuki, R. Tachibana, S. Thomas, B. Ramabhadran, and G. Saon, "Domain adaptation of CNN based acoustic models under limited resource settings," *Proc. Interspeech*, pp. 1588–1592, 2016.

[14] J. Cui, B. Kingsbury, B. Ramabhadran, G. Saon, T. Sercu, K. Audhkhasi, A. Sethy, M. Nussbaum-Thom, and A. Rosenberg, "Knowledge distillation across ensembles of multilingual models for low-resource languages," *Proc. IEEE ICASSP*, pp. 4825–4829, 2017.

[15] T. Fukuda, M. Suzuki, K. Gakuto, J. Cui, S. Thomas, and B. Ramabhadran, "Efficient knowledge distillation from an ensemble of teachers," *Proc. Interspeech*, pp. 3697–3701, 2017.

[16] K. J. Geras, A.-r. Mohamed, R. Caruana, G. Urban, S. Wang, Ö. Aslan, M. Philipose, M. Richardson, and C. A. Sutton, "Compressing LSTMs into CNNs," *CoRR, vol. abs/1511.06433*, 2015.

[17] Y. Chebotar and A. Waters, "Distilling knowledge from ensembles of neural networks for speech recognition," *Proc. Interspeech*, pp. 3439–3443, 2016.

[18] J. Thiemann, N. Ito, and E. Vincent, "DEMAND: a collection of multi-channel recordings of acoustic noise in diverse environments," *Proceedings of Meetings on Acoustics*, 2013.

[19] L. Brandchain, "The mixer 6 corpus: Resource for cross-channel and text independent speaker recognition," *LREC*, 2010.

[20] J. Carletta, "Unleashing the killer corpus: experiences in creating the multi-everything ami meeting corpus," *Language Resources and Evaluation*, vol. 41, no. 1, pp. 181–190, 2007.

[21] S. Itahashi, "Recent speech database projects in japan," *Proc. ICSLP*, 1990.

[22] S. Nakamura, K. Hiyane, F. Asano, T. Nishiura, and T. Yamada, "Acoustical sound database in real environments for sound scene understanding and hands-free speech recognition," *LREC*, 2000.